HUMAN BIOLOGY
a record of research
DECEMBER, 1950
VOL. 22 No. 4

# LONG-TIME FOLLOW-UP IN MORBIDITY STUDIES: THE DEFINITION OF THE GROUP TO BE FOLLOWED

BY PAUL M. DENSEN

*Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh*

FOLLOW-UP studies in relation to morbidity of one kind or another have been appearing in the literature with increasing frequency. This is a direct consequence of the growing importance of the chronic diseases as causes of morbidity and mortality. One cannot describe the natural history of these diseases nor evaluate the efficacy of therapy without being prepared to observe individuals over long periods of time. It is not always understood, however, that the value of these follow-up studies to medical science is proportional to the degree to which there has been precise definition of the groups to be followed. It is the function of the statistician to make this fact clear to the clinician. If he does not do this he is not properly discharging his responsibilities as a consultant. It is for this reason that, although the subject is not a new one, the importance of definition of the group to be followed cannot be stressed too much.

That a clear definition of the group to be followed is a necessary prerequisite to follow-up studies flows from a consideration of the general objective of such studies. Whether the study is concerned with the

natural history of the disease or with the evaluation of therapy, the objective is to obtain a set of facts which may serve as the basis for predictions about similar cases. The cases which are studied are expected to serve as the basis for generalization to this group of similar cases. But it is only to the degree that the study cases have been precisely defined that this similar group can be adequately described and the extent indicated to which generalization may be permitted.

Consider, for example, Doctor X in a hospital who wants to follow cases of, say, "moderate tuberculosis" to evaluate the effectiveness of a certain method of therapy. If he goes to the record room of the hospital and gets only those cases which bear the label "moderate tuberculosis," the results of his study may be of some value to him but they will be of only limited value to the medical profession as a whole. This is because Doctor Y, whose training and experience have been different, might diagnose these same cases as "minimal tuberculosis" or, conversely, might diagnose other cases "moderate" which Doctor X does not consider such at all. In order to get around this difficulty, Doctor X must go behind the label to study the records of the cases in order to define what he means by "moderate tuberculosis." He will probably do this by describing the characteristics of the x-ray film and he will then follow all the cases whose films show this characteristic regardless of whether they are called "moderate tuberculosis" or not. If he does not make this definition from the x-ray film or if he does not make known what definition he has used, Doctor Y may take Doctor X's results and apply them to cases which bear the same label but are not similar to the ones Doctor X used at all.

The foregoing is a fictitious example. It should not be thought, however, that this situation does not exist in actuality. A case in point is afforded by the silicosis problem in the pottery and granite industries. The x-ray film, which is one of the bases for diagnosis of silicosis, looks quite different for the pottery industry from that for the granite industry. In addition, the rate of progression of the x-ray picture is not the same in the two industries. Therefore, if a figure is established for the probability of progression of a certain degree in a certain period of time for the pottery industry it does not necessarily hold for the granite industry. The fact that generalization from one to the other of these industries is unwarranted emerges from the careful description of the cases studied. Unfortunately, in spite of the warnings of authors writing in this field, the findings in these two industries have been generalized to other dust-producing industries.

It may be that for some diseases the criteria for diagnosis are so well established or so objective that there is little variation among the different members of the medical profession in their diagnosis of these cases. Chickenpox or some of the leukemias perhaps might be considered in this class. For such diseases the process of defining the case may be so uniform that there is no need to go behind the label to the record. The number of such "defined" diseases, however, is relatively few. Such diseases as rheumatic fever, whooping cough, tuberculosis, or cancer are certainly not in this category. These are the very diseases to which greatest interest attaches today, and the process of examining the records in order to define the cases must be an integral part of any follow-up program.

Definition from the standpoint of diagnosis is only part of the total problem of definition. Other factors besides diagnosis may play a very important part in the determination of the universe to which the findings may be generalized. Especially is this true when the study is designed to evaluate a particular method of therapy. As an example, one may refer to some preliminary findings in a study, now under way, of aortic insufficiency due to cardiovascular syphilis. Initially, it was thought that it might be possible to evaluate the effectiveness of treatment by using as controls untreated cases obtained from past records of the hospital. The objection was raised, however, that such cases would not be like those of the treated group because there was probably some reason why they were not treated. It was suggested that the records be studied in order to see if this was the case. Tables 1 and 2 are illustrative of the findings.

TABLE 1

*Percentage distribution of cardiovascular syphilitic cases according to evidence of congestive cardiac failure for untreated patients admitted prior to July 1, 1943 and for patients admitted after January 1, 1944*

| EVIDENCE OF CONGESTIVE CARDIAC FAILURE | ADMISSION DATE | |
|---|---|---|
| | Prior to 7/1/43 | After 1/1/44 |
| Total | 100.0 | 100.0 |
| Yes | 53.6 | 35.6 |
| No | 44.3 | 63.4 |
| No record | 2.1 | 1.0 |
| Total cases | 140 | 104 |

TABLE 2

*Percentage distribution of cardiovascular syphilitic cases according to reason for no antisyphilitic treatment for untreated patients admitted prior to July 1, 1943 and for patients admitted after January 1, 1944*

| REASON FOR NO ANTISYPHILITIC TREATMENT | ADMISSION DATE | |
|---|---|---|
| | Prior to 7/1/43 | After 1/1/44 |
| Total | 100.0 | 100.0 |
| Uncooperative | 5.0 | 3.8 |
| Age | 2.9 | 1.9 |
| Death | 0.7 | 1.0 |
| Rx (against medical judgment) | 18.6 | 7.7 |
| Transferred to other hospital or private physician | 4.3 | 8.7 |
| No record | 68.5 | 76.9 |
| Total cases | 140 | 104 |

It is readily seen that the two groups are not the same, either with respect to reason for not receiving treatment or with respect to evidence of congestive failure. Other differences were found, such as differences in age distribution, differences in length of time from initial infection to diagnosis, and so on. As a result of this careful definition of the cases in the proposed control, it was felt that any follow-up study designed to evaluate the efficacy of treatment would lead to unsound generalizations.

This example also illustrates the fact that very often an investigator may have difficulty in finding material to work with of precisely the kind he needs and he may decide to take the best available. This is fair enough as long as he recognizes that the universe to which generalization may then be made may not be the original universe he had in mind. Should the material be found to be selected in some way its use is not necessarily invalidated. The point is that it is essential to specify the nature of the selection in order that others as well as the investigator may make proper use of the findings.

It must be recognized that a definition can be adequate only in relation to some specific objective. For some objectives it may be necessary to engage in very detailed description, such as color of eyes or hair, while for others the fact that an individual is characterized by a certain kind of punch on a punch card may be sufficient. Which of these two extremes is used, or whether a middle ground is sought,

depends entirely upon the universe to which it is desired to generalize the material. If this universe is that of all individuals with a certain punch on a punch card such definition is entirely proper. Usually, however, it is found that if the definition is too broad the results are of no greater value for an understanding of disease problems than a crude death rate is for an understanding of public health problems. On the other hand, too detailed a definition of the group also has its disadvantages.

Mention of definition in the preceding paragraph as a punch on a punch card is deliberate in view of the growing tendency towards co-operative studies and to the use of rosters of cases drawn from punch card files such as those of the Armed Forces or of the Veterans Administration. The large numbers involved in studies drawn from punch card files should not be permitted to obscure the fact that the problem of definition still exists—in fact, it may be multiplied many fold. Consider, for example, the question of "dosage" in infectious hepatitis or malaria or hookworm. The dosages experienced by men in the armed forces in wartime may be very different from those experienced in peacetime situations. Whether the subsequent experience of cases developed during wartime may be fairly generalized to establish the prognosis of cases contracted in other situations can be determined only through adequate definition of the material. This will often require going behind the punch cards to the individual case records, and such examination of the records should be provided for in requests for funds.

In this matter of definition it is the job of the statistician to insist that *a* definition be made and its implications understood. It is not his job, however, to make the definition. This is best left to the individual doing the study—the doctor or the research worker, although such definition may best come out of joint discussion of the problem at its earliest stages. It has been said that "the more thorough the design of the experiment the more meaningful is the question asked." Definition of the group being followed is a part of the design of follow-up studies without which the meaning of the results is ambiguous and, possibly, even dangerous.

# METHODS OF ANALYSIS FOR FOLLOW-UP STUDIES *

BY HAROLD F. DORN

A FREQUENT statistical problem in morbidity studies is the measurement of the rate at which members of a population withdraw from or leave the population. The population may be any defined group of individuals such as persons with cancer who have been given a particular treatment, persons working in a plant where they are exposed to unhealthful conditions, persons admitted to hospital or persons receiving disability benefits when ill.

Members are said to withdraw from the population after the occurrence of some event. Thus for a population of persons who have been treated successfully for cancer, a member may be considered to have withdrawn from the population if he has a recurrence of cancer. In this instance the problem may be to construct a measure of the recurrence rate. Or if the population is composed of persons receiving disability benefits the particular function to be constructed may be one describing the duration of disability.

The measurement of the rate of withdrawal from a population is part of the general theory of decrement functions in population analysis. In the simplest cases, there is only a single decrement function, for example a mortality rate, to be determined. Usually, however, multiple decrement functions must be considered, for even if the only considered reason for withdrawal is recurrence of disease, some members of the population will withdraw due to death from a different disease while others must be withdrawn because they are lost or because of some other event not related to the specific reason for withdrawal under investigation.

The study of populations over extended periods of time is subject to two practical difficulties. The first is the impossibility, without excessive cost, of keeping each member of the population under observation for the duration of the study. Hence only partially complete information will be available for cases lost to follow-up. The second problem arises from the fact that most populations are dynamic, that

* Revision of a paper presented at the Annual Meeting of the American Statistical Association, December 29, 1949.

is, members are continually entering and leaving the group so that at any particular moment of time the population will be composed of persons who have been in it for varying lengths of time. In other words, the members of the population will not have been exposed to the risk of withdrawal for the same length of time. The problem is to combine the experience of persons of different durations is a meaningful way.

Some methods of solving these problems will be discussed in the remainder of this paper. Since only selected aspects of the theory of decrement functions can be discussed, attention will be concentrated on clarifying basic concepts and principles of classifying data. Methods of computation will be mentioned only briefly.

Three concepts underlie the construction of decrement functions, (a) cohort of persons, (b) person-time units of exposure and (c) exposure to risk of withdrawal.

A cohort is a group of persons each of whom has some common characteristic. In follow-up investigations this common characteristic frequently is the fact of entrance into the population under study during the same interval of time. At any anniversary of the date of entry, all members of a cohort will have been in the study for the same duration. The length of time interval of entry is arbitrary and is determined by the conditions of the study.

Members may continually be entering and leaving the population. Even if new persons are not admitted to the study, the original group will be depleted, as time elapses, by death, migration, or some other reason so that the number under observation at the end of the study frequently will differ appreciably from the original number. The risk of withdrawal of a member whether it be by death, recurrence of disease or any other event is a function of the length of time the member is under observation. Consequently, withdrawal rates must be computed per person per unit of time, for example per person per month or per person per year.

The choice of the unit of time for expressing the duration an individual remains in the study is governed by the problem being investigated. If it is a month the numerical unit of the population used in the computation of rates is a person-month, that is, a person observed for a month, and the total experience is the sum of the products of each person by the number of months he is observed. This sum is the number of person-months of exposure.

Although transforming the experience of each member of the popu-

lation to a common time interval such as person-months permits the comparison or combination of the experience of individuals of unequal durations, it does not ensure that the rates of withdrawal based upon the same aggregate number of person-months of exposure are of equal validity. A thousand person-months of exposure may be obtained in many ways, two of which are 1000 persons exposed for one month each and 100 persons exposed for 10 months each. If, as frequently is true, the probability of withdrawal is proportional to the length of time an individual is under observation, careful attention needs to be paid to the distribution of the person-time units of exposure over the total duration of the study.

The computation of the probability of withdrawal for some specific reason depends upon the identification of a cohort of persons each of whom is exposed to the risk of withdrawal for a specified time interval. The probability of withdrawal during a given time interval must be based upon the experience only of those who are exposed to risk. These persons are the exposed-to-risk for that time interval; the number of individuals in the exposed-to-risk group usually will not be the same for all time intervals.

The question to be answered by long-time follow-up studies frequently is, "To what proportion of a group of persons will some specified event happen by the end of a given period of time?" The event may be death, onset of disease, relapse after treatment, recovery from disease or other similar outcomes. The statistical computations required to answer this question appear exceedingly simple. If the proportion of relapses within five years following treatment is wanted, the straightforward thing to do would be to treat some number of persons with disease, count the number of these with a relapse during the subsequent five years and then compute the relapse rate by dividing the number of persons having a relapse by the number originally treated.

This procedure is correct and, if it could be carried out, would yield a satisfactory answer. However, it rarely can be applied. The usual situation is to have a number of different cohorts with varying durations under observation. Moreover some members of a given cohort may have been lost to follow-up, have died from a cause unrelated to the disease being investigated or have been dropped from the study for some other reason. Furthermore, the probability of withdrawal is wanted not for one but for several periods of time. The statistical problem is how to combine heterogeneous data of this nature so that the desired rates can be computed.

The general procedure is to include each individual as long as he is exposed to risk and then to exclude him from the study group. In order to utilize data for persons with unequal duration, the study period is divided into a number of subintervals which need not all be of equal length. Although calculations may be simplified if the subintervals are equal in length, it frequently is advantageous to vary the length so as to make the decrement function approximately linear over the subinterval. The number of individuals exposed to risk at the beginning of each subinterval is computed. These individuals are classified according to their status at the end of the subinterval. Rates are computed based upon the number of persons exposed to risk. These rates are used to construct a stationary population from which the desired probabilities are calculated.

A number of definitions and rules are required. These pertain to (a) cases to be included, (b) the date of entry, (c) the classification of status, (d) the length of the interval of tabulation, (e) the time at which status is determined, (f) the handling of cases observed for part of an interval and (g) the handling of cases lost to follow-up.

The cases to be included must be clearly defined and representative of the population to which generalizations drawn from the observed data are intended to apply. If, as frequently is the case, the data refer to the experience of several cohorts, some check should be made of the homogeneity of this experience.

If the calendar date of initial observation is not the same for all cases, a common time must be chosen in the history of each person from which duration will be measured. Examples of this common initial date on which a person enters the study are the date of diagnosis, the date of initial treatment, the date of completion of treatment, or the date on which a patient is symptom free. The date selected will depend upon the purpose of the study. However, it should be a time which can be determined precisely. The date of completion of treament or the date on which a patient is symptom free have only a limited usefulness in evaluating methods of therapy since persons dying during treatment or cases still uncured after treatment are excluded from the study. This bias would be particularly serious if two methods of therapy such as surgery and radiation or chemotherapy are to be compared.

Categories for classifying the status of each individual at any given time during the study period must be set up. These may include such categories as alive without disease, dead, lost to follow-up and recurrence of disease. It may be desirable to separate the deaths into those resulting

from the disease under study and those due to other causes. Provision also must be made for individuals who are dropped from the exposed-to-risk group because their possible total duration of observation does not extend through the next time interval in the table. For example, if the total study period is three years, individuals whose total possible duration is only eighteen months should be dropped from the exposed-to-risk group at the end of that duration. Criteria for each category should be established in advance so that there will be no uncertainty concerning the classification of any case.

The length of the time intervals into which the entire study period is divided preferably should be equal to the interval between consecutive observations on the individuals in the study. If the outcome is survival or death and the date of death can be ascertained from death certificates or similar records, any appropriate interval for tabulation may be chosen. However, if status must be determined by reexamination of each person, computations will be simplified if the interval between examinations is the same for all persons so that this interval can be used as the interval of tabulation.

Each individual should be classified with respect to his status at the end of each interval of tabulation. Even though a fixed schedule of examination dates is established, the status of every person on these dates usually will not be known due to failure to report for examination. Consequently the dates on which an individual's status is known definitely may not coincide with the end of an interval of tabulation.

For example, suppose that the length of the interval of tabulation is twelve months and the status of an individual is not ascertained until fifteen months after the end of the previous tabulation interval. A problem arises only when status has changed. If the change in status is due to recurrence of disease or to a change in the stage of disease, the only information available frequently will be that the change already had occurred on the date of examination. In such instances, the usual procedure is to assume that the effective date of a change in status is the date on which it is discovered. Thus a person who is alive and free from disease at the beginning of a twelve-month interval and who has a recurrence of disease when next examined after fifteen months is classed as alive and free from disease at the end of the twelve-month interval. This procedure is not ideal but is usually satisfactory if the interval between observations is not too great.

Occasionally status may change more than once within an interval. A person may be alive and free from disease at the beginning of an inter-

val, have a relapse and be successfully retreated so that he is again free from disease at the end of the interval. If the purpose of the study is to determine the proportion of persons who have had a recurrence, these cases would be classed with the recurrent category. But if the objective is to determine the proportion of persons who are alive and free from disease at the end of any specified time interval, these cases would be included in the alive and free from disease category.

The terminal information on individuals whose possible duration of observation is less than the study period rarely will coincide with the end of an interval of tabulation. The individuals exposed to risk at the beginning of an interval are members of two types of cohorts, the first, a cohort for which the end of the period of possible observation falls before the end of the interval and the second, all other cohorts for which the end of the period of possible observations falls at or beyond the end of the interval. A convenient procedure is to drop the former cohort at the end of the last completed interval of tabulation. This discards some information but avoids the trouble of calculating the fractional period of exposure within the incomplete interval. However, it is important to remember that if one member of the cohort is included, all other members also must be included. For example, if a death occurs during the fractional part of the interval it should not be included unless the entire cohort is properly accounted for among the exposed to risk.

If the extra labor is worthwhile, the exact fractional part of the interval during which the individuals are exposed to risk can be computed and the equivalent number of person time intervals of exposure can be determined.

The validity of a follow-up study may be seriously impaired if every person is not observed for the maximum possible duration. It is not possible to set any permissible percentage of persons which may be lost without invalidating the data. Even a small percentage lost to follow-up, less than five per cent of the total number under observation, if this group has a relatively large proportion of persons withdrawing due to the condition being studied, may seriously bias the observations.

The usual situation is to have $E_x$ persons under observation at exact time $x$ and subject to an unknown probability of withdrawal, $\hat{q}_x$, during the succeeding time interval. The expected number of withdrawals is $\hat{q}_x E_x = \hat{W}_x$. The number still in the population at the end of the interval is $E_{x+1}$ and $\hat{q}_x$ is estimated by taking $q_x = W_x/E_x = W_x/(E_{x+1} + W_x)$. Suppose $U_x$ persons are lost to follow-up so that the number actually

observed is $E'_x = E_x - U_x$ and the number observed to withdraw is $W'_x \leqq W_x$. How shall $\hat{q}_x$ be estimated?

It should be emphasized that there is no single correct way of handling persons lost to follow-up. However, three methods merit brief consideration: (a) drop lost persons at the end of the last complete interval of observation; (b) count lost persons as exposed to risk for one-half of the interval immediately following the last complete interval of observation and drop them from all subsequent intervals; and (c) count lost persons both as withdrawals and not as withdrawals in the interval during which they are lost but drop them from all subsequent intervals.

Suppose that the status of each individual is determined by examination at the end of each interval. Persons not reporting for examination are lost to follow-up. Persons lost to follow-up should be dropped on the date of the last examination. The estimate of $\hat{q}_x$ is $q_x = W'_x/E'_x$. This assumes that $(\hat{q}_x U_x)/U_x = (\hat{q}_x E'_x)/E'_x$ where $(\hat{q}_x U_x)$ and $(\hat{q}_x E'_x)$ are the expected number of withdrawals among those who do not and among those who do report for examination respectively. In other words the persons who do not report for examination are assumed to be subject to the same probability of withdrawal as those who report for examination. If this assumption cannot be made there is no correct method of handling the lost persons without additional information.

Suppose it were true that those who are lost came only from those who would withdraw so that the number observed to withdraw is $CW_x$ where $C$ is a fraction. The observed probability of withdrawal would be $CW_x/(E_{x+1} + CW_x)$ which is less than $W_x/E_x$. Hence some quantity must be *subtracted* from $(E_{x+1} + CW_x)$ to obtain an unbiased estimate of $\hat{q}_x$.

On the other hand, suppose that the lost to follow-up come only from $E_{x+1}$ so that the number observed to be still in the population at the end of the interval is $CE_{x+1}$. Then the observed probability of withdrawal would be $W_x/(CE_{x+1} + W_x)$ which is greater than $W_x/E_x$. Hence some quantity must be *added* to $(CE_{x+1} + W_x)$ to obtain an unbiased estimate of $\hat{q}_x$.

In instances when the procedure of dropping lost cases at the end of the last completed interval of observation cannot be used, it is customary to assume that the situation described in the preceding paragraph exists, namely that some quantity should be added to the denominator of the ratio to obtain an unbiased estimate of $\hat{q}_x$. The estimate of $\hat{q}_x$ is taken as $q_x = W'_x/(E_{x+1} + W'_x + \frac{1}{2}U_x)$. This assumes that $W'_x = \hat{q}_x(E_{x+1} + W'_x + \frac{1}{2}U_x)$ so that the observed number of withdrawals is the sum of

withdrawals arising from $\frac{1}{2}U_x$ and withdrawals arising from the group kept under observation. This is equivalent to assuming that the lost to follow-up drop out uniformly throughout the interval so that the number dropping out during any fraction, $\Delta t$, of the interval is $\Delta t U_x$. Moreover, $U_x$ is assumed to be unbiased with respect to the probability of withdrawal for if $q_x$ is an unbiased estimate of $\hat{q}_x$ then $(\hat{q}_x \frac{1}{2}U_x)/\frac{1}{2}U_x$, which of course is unknown, must also be an unbiased estimate of $\hat{q}_x$.

The procedure just described almost certainly is not applicable in many instances when it is used. Moreover it is the wrong kind of correction if a larger proportion of $W_x$ than of $E_{x+1}$ are lost. The only correct method of handling persons lost to follow-up is not to have any.

However, lost persons frequently cannot be prevented without excessive cost so that they must be disposed of in some way in the calculations. If the assumptions discussed above cannot be made, the third procedure previously mentioned can always be used, that is, the persons lost to follow-up are counted both as withdrawing and as not withdrawing in the interval during which they are lost and are dropped from the population exposed to risk at the end of the interval. Considering all the $U_x$ as withdrawals gives a maximum withdrawal rate, $(W_x + U_x)/E_x$, while considering none of the $U_x$ as withdrawals gives a minimum withdrawal rate, $W_x/E_x$. The true withdrawal rate will lie between the maximum and minimum rates. Moreover, the maximum withdrawal rate is equal to the sum of the minimum withdrawal rate and the proportion lost to follow-up.

This procedure always is correct and has the advantage of clearly showing the effect of lost cases upon the conclusions. Its disadvantages are that it increases the labor of computation and does not yield a unique withdrawal rate. However, if persons are lost to follow-up there is no known correct withdrawal rate so that perhaps it is an advantage to show a range within which the correct rate must fall.

The procedure for computing withdrawal rates depends upon the type of decrement table. Decrement tables may be grouped into (a) single decrement tables involving either a single or two or more cohorts and (b) multiple decrement tables involving either a single or two or more cohorts.

In a closed population of $E$ individuals all belonging to a single cohort and with a single cause of withdrawal, for example death, if $W_x$ individuals withdraw during the interval $x$ to $x + 1$ out of $E_x$ individuals exposed to risk at exact time $x$ the probability of withdrawal is $W_x/E_x$. This is true whether the interval of time is a calendar year of age or

some other interval measured from an arbitrary date of entry into the population, for example the date of diagnosis of disease. The essential requirement is that both $E_x$ and $W_x$ must relate to the same group of persons, that is, all persons exposed to risk of withdrawal for the entire interval. Furthermore even though the persons who withdraw do so at various times throughout the interval they must be considered as exposed to the risk of withdrawal for the entire interval; otherwise the probability of withdrawal could be greater than unity. This also may be seen from the fact that $W_x/E_x$ may be written as $W_x/(E_{x+1} + W_x)$ where $E_{x+1}$ is the number still in the population at the beginning of the next interval and consequently exposed for the entire interval, $x$ to $x + 1$. These definitions hold for populations composed of several cohorts provided $E_x$ and $W_x$ are pooled for identical intervals since the date of entry of each cohort.

If the population is subject to multiple causes of decrement, the calculations become somewhat more complicated. For simplicity only a population composed of a single cohort and subject to two causes of withdrawal, $W_1$ and $W_2$, will be considered. $W_1$ might be recurrence of disease and $W_2$ might be death before recurrence of disease.

Suppose we wish to measure the combined effect of the two causes of withdrawal. Since a person must (a) withdraw due to $W_1$, (b) withdraw due to $W_2$ or (c) remain in the population, we can write $(W_{1x}/E_x) + (W_{2x}/E_x) + (E_x - W_{1x} - W_{2x})/E_x = 1$ and the separate terms are respectively, the probability of withdrawing due to $W_1$, the probability of withdrawing due to $W_2$ and the probability of not withdrawing. It is apparent that these are conditional probabilities for as $W_{2x}/E_x$ approaches one, $W_{1x}E_x$ and $(E_x - W_{1x} - W_{2x})/E_x$ must approach zero. Nevertheless this formula is suitable for expressing the combined effect of the multiple causes of withdrawal.

Assume that it is desired to express the probability of withdrawal due to $W_1$ independent of $W_2$. In general, this cannot be done unless $W_1$ and $W_2$ are actually independent; however it is possible to express the probability of withdrawal due to $W_1$ among persons who do not withdraw due to $W_2$.

Let $W_{1x}/E_x = q_{1x}$. As pointed out above this is the conditional probability of withdrawal due to $W_1$ during the interval $x$ to $x + 1$. $E_x$ may be written as $E_{x+1} + W_{1x} + W_{2x}$ showing that it includes persons withdrawing due to $W_2$. Obviously the $W_{2x}$ could not be exposed to the probability of withdrawal due to $W_1$ after they withdrew from the population.

The correct procedure is to include $W_{2x}$ for the portion of the interval, $x$ to $x+1$, prior to withdrawal and then to exclude them from $E_x$. The general principle is that $N$ persons each exposed for $1/K$ part of an interval are equivalent to $N/K$ individuals exposed for the entire interval. If the principle is used, withdrawals should be spread fairly uniformly over the interval, otherwise the withdrawal rate due to $W_1$ may be biased.

If the exact date of withdrawal of $W_{2x}$ is known, the exact number of equivalent person-time units of exposure can be calculated. However, if it can be assumed that the number of withdrawals during any fraction, $t$, of the interval, $x$ to $x+1$, is $t$ $W_{2x}$ we have $\int_0^1 E_{x+t}dt = E_x - \frac{1}{2}W_{2x}$ as the equivalent number of persons exposed to the risk of withdrawal due to $W_1$. Hence we may take $r_{1x} = W_{1x}/(E_x - \frac{1}{2}W_{2x})$ as the probability of withdrawal due to $W_1$ among persons not withdrawing due to $W_2$. Correspondingly we have $r_{2x} = W_{2x}/(E_x - \frac{1}{2}W_{1x})$. It is important to note that there is no simple quantity which added to $r_{1x}$ and $r_{2x}$ will equal unity as was the case with $q_{1x}$ and $q_{2x}$.

The simplest procedure is first to compute $q_{1x} = W_{1x}/E_x$ and $q_{2x} = W_{2x}/E_x$ and then to construct a stationary population from these values by choosing some convenient initial number or radix, $l_0$, say 1,000 or 100,000. Then $w_{10} = l_0 q_{10}$, $w_{20} = l_0 q_{20}$ and $l_1 = l_0(1 - q_{10} - q_{20})$. By repeated application of these formulas, the entire table can be constructed.

Then $q_{1x} = w_{1x}/l_x$ will be the probability that a person in the population at the beginning of the $(x+1)$st interval will withdraw due to $W_1$ during the interval. The corresponding probability for $W_2$ will be $q_{2x} = w_{2x}/l_x$ and $q_{1x} + q_{2x} + (l_{x+1}/l_x) = 1$.

Also $r_{1x} = w_{1x}/(l_x - \frac{1}{2}w_{2x})$ is the probability that a person in the population at the beginning of the $(x+1)$st interval will withdraw due to $W_1$ prior to withdrawal due to $W_2$ during the interval. Or in other words, it is the probability of withdrawal due to $W_1$ among persons who do not withdraw due to $W_2$. The corresponding probability for $W_2$ is $r_{2x} = w_{2x}/(l_x - \frac{1}{2}w_{1x})$.

It is apparent that $r_{1x} \geqq q_{1x}$. It can be shown that $q_{1x} = r_{1x}(1 - \frac{1}{2}q_{2x})$ from which $r_{1x} = 2q_{1x}(2 - q_{2x})^{-1}$. The two values of $r$ are related by the formula $r_{1x} = q_{1x}[1 - \frac{1}{2}r_{2x}(1 - \frac{1}{2}q_{1x})]^{-1}$.

The values of $r$ are related to the central withdrawal rates in the same way that $q_x$ and $m_x$ are related in a mortality table which is, of course, a single decrement table. Using the assumption that the number of

withdrawals in the stationary population during any fractional part, $t$, of the interval $x$ to $x+1$ is $tw_x$, the average number of persons exposed to risk, $L_x$, would be $\int_0^1 l_{x+t}dt = l_x - \frac{1}{2}(w_{1x} + w_{2x})$.

The central withdrawal rate due to $W_1$ would be $m_{1x} = w_{1x}/L_x$ and since $r_{1x} = w_{1x}/(L_x + \frac{1}{2}w_{1x})$ we have $r_{1x} = 2m_{1x}/(2 + m_{1x})$. Similarly $r_{2x} = 2m_{2x}/(2 + m_{2x})$.

# TIMING OF THE DISTRIBUTION OF EVENTS BETWEEN OBSERVATIONS*

## A CONTRIBUTION TO THE THEORY OF FOLLOW-UP STUDIES

BY

T. E. HARRIS

*RAND Corporation*

PAUL MEIER

*Philadelphia Tuberculosis and Health Association*

AND

JOHN W. TUKEY

*Princeton University*

### 1. *Introduction and Summary*

WHEN all that is known as to the time of occurrence of an event is that it had not occurred at one known time ("last previous") and that it had occurred at another known time ("first after"), it is ordinarily useless to try to time the event more closely. But when we have information of this nature about many events, and when it is reasonable to think of the events as a sample from a distribution, it is possible to use the information to estimate the distribution. It is natural to express this distribution in terms of event-rates. The problem discussed here is how such event-rates may reasonably be estimated from such observations. We are trying to learn as much as we can from case histories with serious gaps.

Any attempt to infer the timing of events from sparse observations is subject to many pitfalls. Some of these are discussed briefly in the first part of the paper.

If a given body of data has avoided these pitfalls, a simple statistical model may apply, in which the standard estimation procedures, such as

* Prepared in connection with research sponsored by the Office of Naval Research.

maximum likelihood, are available. In the third part of the paper it is shown that maximum likelihood estimates, which some statisticians regard as best possible estimates, are closely approximated by those based on a very simple procedure. This procedure is the following:

(1) Each case between its "last previous" and "first after" contributes $\frac{1}{2}$ to the number of cases exposed to risk;

(2) Each case's event is distributed over the interval between "last previous" and "first after" according to the finally estimated chance of occurrence.

We emphasize that this "$\frac{1}{2}$-rule" is not an assumption, but the natural, mathematical, first approximation to maximum likelihood in a situation where even a second approximation is rarely needed.

An iterative procedure of obtaining exact maximum likelihood estimates, and making rough (but usually adequate) significance tests follows in the last section.

The second part of the paper illustrates the results and proposes modifications. A small example of the simplified procedure is given in Section 7. This is followed by a discussion of the danger that maximum-likelihood-like estimates may follow the irregularities of the data too faithfully, and the presentation of some adjusted procedures. Then follow some examples of the results of applying the full maximum likelihood procedure to the "progression" of cases of minimal tuberculosis. These are based on the records of the Henry Phipps Institute as studied by Drs. A. L. Cochrane, H. W. Campbell and S. C. Stein [2]. The present analysis confirms their conclusions, as obtained by simpler methods, and exhibits a marked tendency for the progression dates to avoid a period of 9 to 12 months after detection. The possible significance of this result is discussed, and it is concluded that modified follow-up schedules would have to be applied to many new cases before its reality could be established.

## I. THE PROBLEM

### 2. *The basic situation*

A certain number of cases are examined at irregular intervals, varying from case to case, to determine whether a certain event has occurred. For each examination the result is merely "event has not yet occurred" or "event has occurred." Cases are "lost" from the record for various reasons, one being arrival at date of termination of study before the event occurs, so that the only thing known about them is that the event had

not occurred up to a certain date. These are "lost" cases, in contrast to the "eventful" cases where the event is known to have happened between certain dates.

The particular situation which led to our study of this problem was the medical one used as an example in Section 8. The cases were individuals whose chest x-rays were read at the Henry Phipps Institute as (active, questionable or inactive) minimal tuberculosis during the years 1925 to 1945. The observations were the diagnoses that the tuberculosis had not or had become advanced (= had progressed). The date was the number of months elapsed from detection to progression. (Notice that the event is assumed to be reaching a state where examination ensures a diagnosis of progression. We are leaving aside any discussion of the accuracy of diagnosis.)

The intervals at which individuals were reexamined were irregular, and many cases were lost from the records after a relatively small number of months. Others are lost cases because they have not yet progressed, or because they have died from other causes. All these facts can lead to pitfalls.

Similar situations arise frequently in clinical follow-ups of all sorts, and less frequently in diverse fields.

3. *Losses*

In order to analyze data of the sort just described, we must make some assumption about the losses—about those cases which disappear from the record before the event occurs. There are three kinds of such cases, those really lost during the period of the study, those dying from other causes, and those terminated by closing of the study. In some situations it may be possible to follow-up a small sample of the really lost cases in order to learn how they compare with those not lost. Such studies can be very helpful in leading to an intelligent assumption about such losses. Follow-up of cases lost by death is obviously impossible, and differences in the character of terminated cases form part of the ever present problem of time trend. In the absence of follow-up data on a sample of lost cases or other special information, the assumption will usually be made that the lost cases are so similar to those not lost that the eventful cases are representative. We need now to discuss the errors in such a procedure.

An interesting point in clinical follow-up is the treatment of deaths presumed to be unrelated to the event. If susceptibilities to the event

and to death from other causes are correlated, then the cases lost through death, if they happened to live on, would have been more or less susceptible to the event, at dates after their actual deaths, than would those actually exposed. For many purposes it is most convenient to avoid this question by stating firmly that our interest is in the distribution of events among those not dying from other causes. We shall, indeed, adopt this position in our analyses. We must remember, however, that changes in the death rates from other causes may, and in case of correlated susceptibility, must, affect our event rates.

Those lost from the record from other causes than death may be reasonably divided into those moving out of the geographical or other area covered, and others. It would be possible to deal with emigrants as we have just proposed to deal with deaths—but the justification is much weaker. The procedure of defining them out of existence is only reasonable if the event-rate is widely different in the survey area as compared to the new locations of emigrants. Since emigration rates are usually much less stable than death rates, their possible changes are more threatening.

Finally the mysterious group of "others." In a clinical follow-up situation, we always have the suspicion that this group is quite different from the group of cases which do not disappear without explanation. But we rarely know to what extent this is true.

When the group of cases falls into natural categories, such as sex, color, age, date of entry, economic status, weight, etc., it may be profitable to estimate event-rates by category and combine according to the original number of cases. This allows for any tendency of the lost cases to be unequally distributed among the categories.

Notice that we are always interested here in event-rates for the population represented by the group of cases initally considered. There is no assumption that this population is homogeneous.

With these pitfalls in mind, the desirability of some analysis leads us to assume (possibly in desperation) that systematic differences between lost cases and eventful cases can be neglected. Our mathematical models will neglect any such difference.

### 4. *Event-observation relations*

Even if there were no lost cases, and our records went on and on for every case, there would remain a closely related source of pitfalls. If the pattern of observations on an individual case depends on, *or is related to* the occurrence of the event, then we need to know in what way, and how

much, in order to analyze the record properly. Thus, if observation requires return to a clinic, the effects on the record of (a) immediate return after the event, or (b) 4-month delay in return after the event, are quite different. These problems are harder to study than those connected with losses. About all we can do is to point out their existence and say that we shall have to neglect them. The most that we shall do is to point out the existence of subtler pitfalls of this sort. If, for example, one group of cases tend to be observed after 0, 4, 12, 16, 24, 28, 36, . . . months, and another group tends to be observed after 0, 8, 12, 20, 24, 32, 36, . . . months, then differences in event-rates in the groups could affect the analysis, even if there were no causal connection between date of event and date of next observation because there will be a (probability) relation between the occurrence of the event and the pattern of observations.

5. *Model*

With these pitfalls noted, we shall go on as if they were not there. We want a specific model on the basis of which we can assess event-rates, and discuss their accuracy. As usual, this model will take the form of a statement that certain things which did not happen are equally probable with what did.

We shall suppose that there is a population, whether large and finite or infinite will not matter, of potential cases falling in the category we are analyzing. Each of these potential cases has the event occur at a definite date, or not at all, as the case may be. There are a group of $N$ patterns of observations, one of which might be 0, 3, 9, 27, 29, 35, 57, 65 (lost after this date) months. The patterns are, in fact, just those in the record. A sample of $N$ potential cases are randomly selected from the population of potential cases, and randomly assigned to the $N$ patterns of observation. These random processes of selection and assignment define the other equally likely records with which the actual record is to be compared.

Some of the ways in which this model is idealized are clear from the previous discussion. It has, however, two advantages:

(1) it is definite; and

(2) it does not require a homogeneous population.

Let us now write $Q_n$ for the fraction of potential cases in which the event does not occur before date $n$. Then it is easy to see that a case recorded

up to $n$ months has probability $Q_n$ of being a lost case. Conversely, the probability that a case will be last seen before at date $n$ and first seen after at date $n + m$ is

$$Q_n - Q_{n+m}.$$

The consistency of these probabilities may be seen from

$$1 \equiv (1 - Q_{t_1}) + (Q_{t_1} - Q_{t_2}) + \cdots + (Q_{t_{k-1}} - Q_{t_k}) + Q_{t_k},$$

where $t_1, t_2, \cdots, t_k$ are the dates on which this case was actually observed.

If we wish to work with a continuously changing event-rate $\rho(t)$, we have

$$Q_n = e^{-\int_0^n \rho(t)\,dt}.$$

Various special cases are discussed in Part III.

6. *Comparison with other approaches*

The problem we attack here differs in two ways from that discussed by Dorn [3] in the preceding paper. First, we regard wide gaps and gross irregularities of pattern in the records of the individual cases as the usual thing. Dorn discusses allowances for a small number of such difficulties, and these allowances seem adequate when regularly-spaced examinations are the usual thing. Unfortunately, this is not always the case.

Second, Dorn seeks probabilities of occurence in a given interval, while we seek rates. A rate of 20 per cent per month for ten months is entirely reasonable and corresponds to a probability of occurence

$$1 - e^{-(200\%)} = 1 - e^{-2} = .865 = 86.5\%.$$

The fact that we seek rates rather than probabilities for intervals produces certain differences in our treatment of observations extending over parts of intervals as compared to the "actuarial" treatment. Each is appropriate in its place.

Various forms of a "$\frac{1}{2}$-rule" have been used or suggested from time to time, for example, by Greenwood [4], but we believe that the present treatment is the first in which it is derived as a natural approximation to an estimate based on general theoretical principles.

Our methods are appropriate for the early investigation of an event-rate. Once much is known about a particular problem, it often becomes

possible to assume a simple functional form involving a few constants for the event-rate as a function of the time. One case, where gaps between observations were no problem, has been treated by Boag [1]. There is no theoretical difficulty in extending such a treatment (for example, along the lines of the last section of the present paper) to the case of serious gaps.

## II. EXAMPLES

### 7. *A simple example of the simple method*

The analysis in Part III leads to the suggestion of the following simple method:

(1) Break the time scale into intervals and consider the event-rate constant in each. Call the value in the $i$-th interval $\rho_i$.

(2) Treat each case to which the event is known not to have occurred as one case exposed to risk, and each case between the dates last before and first after as one-half case exposed to risk.

(3) Prorate the occurrence of an event over the intervals in which it might have occurred according to the product of length of exposure and estimated event-rate.

The last prescription leaves us with an iterative process to obtain the solution.

Let us introduce

$$n_i = \text{time in interval } i \text{ before date last before or date of loss,}$$

$$m_i = \text{time in interval } i \text{ between date last before and date first after.}$$

Then the prorated fraction of an event in interval $i$ is

$$f_i = \frac{\rho_i m_i}{\rho_1 m_1 + \rho_2 m_2 + \cdots + \rho_k m_k}.$$

The procedure is best understood by analyzing a simple example of unrealistic smallness. The essentials of the record are given in Table 1.

If we decide to analyze the example in intervals of length 5, we can find the $n_i$ and $m_i$ with the results in Table 2.

TABLE 1

*Synthetic illustrative example*

| CASE NUMBER | DATE LAST PREVIOUS | DATE FIRST AFTER | REMARKS |
|---|---|---|---|
| 1 | 10 | 16 | .. |
| 2 | 2 | 6 | .. |
| 3 | 3 | 4 | .. |
| 4 | 14 | 15 | .. |
| 5 | 20 | .. | No further record |
| 6 | 25 | .. | Deceased 25 |
| 7 | 30 | .. | Currently at date 31 |
| 8 | 12 | 20 | .. |

TABLE 2

*Values of $n_i$ and $m_i$ for intervals of length 5*

| CASE | $n_1$ | $n_2$ | $n_3$ | $n_4$ | $n_5$ | $n_6$ | $m_1$ | $m_2$ | $m_3$ | $m_4$ | $m_5$ | $m_6$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 5 | 5 | . | . | . | . | . | . | 5 | 1 | . | . |
| 2 | 2 | . | . | . | . | . | 3 | 1 | . | . | . | . |
| 3 | 3 | . | . | . | . | . | 1 | . | . | . | . | . |
| 4 | 5 | 5 | 4 | . | . | . | . | . | 1 | . | . | . |
| 5 | 5 | 5 | 5 | 5 | . | . | . | . | . | . | . | . |
| 6 | 5 | 5 | 5 | 5 | 5 | . | . | . | . | . | . | . |
| 7 | 5 | 5 | 5 | 5 | 5 | 5 | . | . | . | . | . | . |
| 8 | 5 | 5 | 2 | . | . | . | . | . | 3 | 5 | . | . |
| Total | 35 | 30 | 21 | 15 | 10 | 5 | 4 | 1 | 9 | 6 | 0 | 0 |
| Exposure | 37 | $30\frac{1}{2}$ | $25\frac{1}{2}$ | 18 | 10 | 5 | | | | | | |

Thus, for example, Case 8 was last before at date 12, contributing $n_1 = 5$, $n_2 = 5$, $n_3 = 2$ for the 12 units up to this date. The same case was first after at 20, and was thus doubtful for the intervals 12-15, 15-20, which contribute $m_3 = 3$, $m_4 = 5$. All other values of $n_i$ and $m_i$ for this case are zero, and are shown as dots. The exposures in the last line are equal to the total $n_i$ plus one-half the corresponding $m_i$. Thus, $25\frac{1}{2} = 21 + \frac{1}{2}(9)$.

We now have the effective number of case-units exposed to risk in each interval. What remains is to prorate the occurrence of the events. This is not a problem for cases 5, 6, and 7 for which no event is recorded, or for cases 3 and 4 where the event must have occurred in intervals 1 and 3 respectively. The iterations will thus involve only the prorating

of the events in cases 1, 2, and 8. These proratings will not affect intervals 5 and 6, and so we find the estimates

$$\rho_5 = \frac{\Sigma f_5}{\Sigma n_5 + \frac{1}{2}\Sigma m_5} = \frac{0}{10} = 0,$$

$$\rho_6 = \frac{\Sigma f_6}{\Sigma n_6 + \frac{1}{2}\Sigma m_6} = \frac{0}{5} = 0.$$

Now we can begin by supposing $\rho_1 = \rho_2 = \alpha$, (the value of $\alpha$ will not matter) when we prorate the event in Case 2 according to

$$f_1 = \frac{3\alpha}{3\alpha + \alpha} = 0.75,$$

$$f_2 = \frac{\alpha}{3\alpha + \alpha} = 0.25.$$

The result is a contribution of $\frac{3}{4}$ of an event in interval 1 and $\frac{1}{4}$ of an event in interval 2. Case 3 contributes 1 event to interval 1. The other cases contribute nothing. Hence, the first approximations are

$$\rho_1 = \frac{1.75}{37} = 0.0473 = 4.73\%,$$

$$\rho_2 = \frac{0.25}{30\frac{1}{2}} = 0.0082 = 0.82\%.$$

This concludes the first cycle.

Now we find for Case 2 that

$$\rho_1 m_1 = 3(4.73\%) = 14.19\%,$$

$$\rho_2 m_2 = 1(0.82\%) = 0.82\%,$$

so that the prorating factors are

$$f_1 = \frac{14.19}{14.19 + 0.82} = .95,$$

$$f_2 = \frac{0.82}{14.19 + 0.82} = .05.$$

Thus, the second approximations to our estimates are

$$\rho_1 = \frac{1.95}{37} = 0.0528 = 5.28\%$$

$$\rho_2 = \frac{0.05}{30\frac{1}{2}} = 0.0016 = 0.16\%.$$

This is the end of the second cycle.

Now we repeat again to find

$$f_1 = \frac{15.84}{15.84 + 0.16} = .99,$$

$$f_2 = \frac{0.16}{15.84 + 0.16} = .01,$$

$$\rho_1 = \frac{1.99}{37} = 0.0538 = 5.38\%,$$

$$\rho_2 = \frac{0.01}{30\frac{1}{2}} = 0.0003 = 0.03\%.$$

We now see that the eventual result will be

$$\rho_1 = 5.4\%,$$

$$\rho_2 = 0.0\%.$$

Because this was an oversimplified example we could iterate $\rho_1$ and $\rho_2$ without attention to $\rho_3$ and $\rho_4$. In a realistic example, all $\rho_i$ would almost certainly have to be iterated together. We can here, however, iterate $\rho_3$ and $\rho_4$ separately also, and the results are shown in Table 3.

TABLE 3

*Successive iterations toward event rate estimates for the example of Table 1*

(Expressed in per cent per time unit)

| ITERATION | $\rho_1$ | $\rho_2$ | $\rho_3$ | $\rho_4$ | $\rho_5$ | $\rho_6$ |
|---|---|---|---|---|---|---|
| Start | $\alpha$ | $\alpha$ | $\beta$ | $\beta$ | $\gamma$ | $\delta$ |
| First | 4.7 | 0.8 | 8.7 | 4.4 | 0.0 | 0.0 |
| Second | 5.3 | 0.2 | 9.6 | 3.1 | | |
| Third | 5.4 | 0.0 | 10.2 | 2.3 | | |
| Ultimate | 5.4 | 0.0 | 11.7 | 0.0 | 0.0 | 0.0 |

Whether we attach any significance to the estimates in the last line or not, when so small a sample as this is involved, these simple estimates are shown in Part III to closely resemble the maximum likelihood estimates based on the same observations.

8. *Discussion of estimates*

The estimates just obtained in the simple example, and their maximum likelihood cousins, show a pronounced failing—a tendency to follow the details of the observations too far. In the example, the pro-

rating of all three events which might have fallen in one of two intervals was finally of the form 100 per cent for one, 0 per cent for the other. When larger bodies of data are analyzed, the effects will not be so striking, and may indeed be negligible, but the tendency will lie in the same direction.

This is a general tendency of maximum likelihood estimates, and their relatives. Any estimating process which prorates incompletely classified observations, weights observations, or judges the probability of observations, on the basis of the final estimates inevitably shows this tendency. The estimate will always be biased—biased toward roughness in this case, toward small variances in fitting location and scale of a normal distribution, toward distorted values in fitting variance components, and so on.

Thus, in critical cases, we may reasonably take some smooth starting values, and iterate only once, or perhaps twice. The values at this stage may be the safest to use. In the simple example, we may choose to accept

$$\begin{aligned}\rho_1 &= 4.7,\\ \rho_2 &= 0.8,\\ \rho_3 &= 8.7,\\ \rho_4 &= 4.5,\\ \rho_5 &= 0.0,\\ \rho_6 &= 0.0,\end{aligned}$$

as the best values the very limited record can provide, and suspect that $\rho_1 = \rho_2 = \rho_3 = \rho_4 = \rho_5 = \rho_6 = 4$ might be adequate.

9. *The minimal tuberculosis example*

The sources and handling of this data have been described by Cochrane, Campbell and Stein [2]. For the present purpose we note only that:

(1) The dates were given to the nearest month;

(2) The basic categories were active (code 81), questionable (code 84), and inactive (code 87) minimal tuberculosis;

(3) Other information was available as to sex, color, etc.

The fitting was actually performed by the method outlined in Section 11 rather than by the simple method.

For the pooled minimals, the fitted values of $\rho$ and the intervals concerned were as shown in Table 4. (More than 600 of the original 1100 cases were being followed after 24 months.)

TABLE 4

*Progression-rates for pooled minimal tuberculosis*

(Expressed in per cent per month)

| INTERVAL NUMBER | DATES COVERED | ESTIMATED PROGRESSION-RATE |
|---|---|---|
| 1 | 0-3 | 1.33 |
| 2 | 3-6 | 0.95 |
| 3 | 6-9 | 0.63 |
| 4 | 9-12 | 0.01 |
| 5 | 12-15 | 0.65 |
| 6 | 15-18 | 0.63 |
| 7 | 18-21 | 0.89 |
| 8 | 21-24 | 0.15 |
| 9 | 24- | 0.11 |

If the true progression-rates follow the course of the estimated progression-rates, then Table 4 is evidence pointing toward two distinct periods of progression, one under 9 months, and the other over 12 months. There are three possibilities:

(a) The apparent division into two parts is an artefact,—is due to sampling fluctuations and susceptibility of analytic procedure;

(b) There are two peaks, but they are due to obvious factors such as sex, color, age, apparent severity of disease, etc.;

(c) There are two peaks, and they represent two patterns of disease—patterns which may appear in any sort of individual.

We shall not try to settle this question here, but we shall provide some further relevant evidence.

In order to settle this question, it will almost certainly be necessary for some clinic to lay considerable emphasis on examination at *both* 8-9 months *and* 12-13 months after diagnosis as minimal tuberculosis. Attempts to make statistical inferences from a greater quantity of the sort of observations now available will be difficult, probably requiring impractical amounts of data.

Since differences in apparent severity of the disease at diagnosis seemed likely to be reflected in differences in time of progression, the first

break-down was made according to active (code 81) and questionable (code 84). For convenient and effective analysis it was assumed that the progression-rate for a class was the product of the Table 4 rate and a multiplier. The multipliers were then estimated by iterating to the least square solution. The estimated multipliers, shown in Table 5, give no

TABLE 5

*Estimated multipliers to be applied to Table 4 progression-rates*

| PERIOD | ESTIMATED MULTIPLIER | |
|---|---|---|
| | Active | Questionable |
| 0-8 | $2.4 \pm 0.4$ | $0.50 \pm 0.14$ |
| 8-20 | $3.3 \pm 0.7$ | $0.40 \pm 0.18$ |
| 20- | $2.8 \pm 0.7$ | $0.70 \pm 0.20$ |

indication that apparent severity is related to the pattern of progession-rate. (Since only 6 cases classified "inactive" progressed, no computations were made for this group.)

An investigation of the effects

(1) sex and color,

(2) age at time of detection, and

(3) economic status,

was carried out. For example, the expected number of progressions in each month was calculated for each sex, color group. These were adjusted by a separate factor for each group to make total expected equal to the total observed, and the agreement tested by chi-square. Although the overall rates differed from group to group, in no instance was the observed pattern significantly different from that for the entire group.

The difficulty in judging the significance of the "dip" lies in what we shall use as a standard of comparison. What sort of event-rate *vs.* time curve shall we fit, and what sort of deviation from it shall we test? This is a familiar dilemma when trying to study critically a hypothesis suggested by the data, and we can contribute little to the general problem, or to this particular case. The most illuminating information we can supply is given in Table 6. The event-rates in Table 4 were actually

TABLE 6

*Indicated sizes and standard deviations of the splitting parameters for the minimal tuberculosis example*

| PERIOD | ORIGINAL EVENT-RATE | HALF-DIFFERENCE OF EVENT-RATES | |
|---|---|---|---|
| | | Estimate | Std. deviation |
| 0-6 | 1.14 | +.19 | .26 |
| 6-12 | 0.32 | +.31 | .20 |
| 12-18 | 0.64 | +.01 | .30 |
| 18-24 | 0.52 | —.37 | .33 |

obtained by the following process:

(1) The iterative process was carried out for double-length periods of 0-6, 6-12, 12-18, 18-24, 24- months, enough iterations being made to give reasonably stable values;

(2) Each interval was separately split in half, and the constant event-rate found by iteration for a period was replaced by two event-rates in the two half periods, whose average was forced to equal the original iterated overall rate;

(3) The magnitudes of the shifts, and the corresponding standard deviations were estimated as in Section 11.

Thus, these values are not the results of final iteration—but, as pointed out in Section 8, this may not be a disadvantage.

For the 6-12 month period the change is about 1.5 standard deviations. The others are smaller. On the basis of this comparison alone we can only report the dip as a possible trend requiring more data for confirmation or denial.

## III. THEORY

### 10. *Case of constant* $\rho$

This case is oversimplified for most interesting problems, but by this oversimplification, we can see quite a lot of the essentials through a relatively thin skin of notation.

The observational record, just as in all other cases, will consist of lists of

$$n = \text{date of loss (lost cases)}$$

$$\left.\begin{aligned} n &= \text{last date before} \\ n+m &= \text{first date after} \end{aligned}\right\} \quad \text{(eventful cases).}$$

The contribution of a lost case to the likelihood is just the probability of a random case reaching date $n$ without event, namely

$$e^{-\rho n}.$$

The contribution of an eventful case to the likelihood is the joint probability of a random case reaching time $n$ without event and then having an event before time $n+m$. But the desired probability is just the difference between the probabilities of no event up to the time $n$ and up to time $n+m$. Hence, the contribution to the likelihood is

$$e^{-\rho n} - e^{-\rho(n+m)} = e^{-\rho n}(1 - e^{-\rho m}).$$

Thus, every case contributes a factor $e^{-\rho n}$ to the likelihood, and each eventful case contributes an additional factor $(1 - e^{-\rho m})$. Thus, the logarithm of the likelihood is given by

$$\Sigma(-\rho n) + \Sigma' \log(1 - e^{-\rho m}),$$

where, as hereafter, $\Sigma$ indicates summation over all cases, and $\Sigma'$ indicates summation over eventful cases.

The derivative with respect to $\rho$ is

$$-\Sigma n + \Sigma' \frac{m e^{-\rho m}}{1 - e^{-\rho m}} = \frac{1}{\rho}\left(-\Sigma n + \Sigma' \frac{\rho m}{e^{\rho m} - 1}\right),$$

and the maximum likelihood estimate of $\rho$ is found by equating this to zero.

Now

$$\frac{u}{e^u - 1} = 1 - \tfrac{1}{2}u + \frac{u^2}{12} - \frac{u^4}{720} + \text{higher terms},$$

so that

$$\Sigma' \frac{\rho m}{e^{\rho m} - 1} = \Sigma' 1 - \rho \Sigma' \frac{m}{2} + \frac{\rho^2}{12} \Sigma' m^2 - \frac{\rho^4}{720} \Sigma' m^4 + \text{higher terms},$$

and the maximum likelihood estimate comes from

$$\frac{1}{\rho}\left(-\rho \Sigma n + \Sigma' 1 - \rho \Sigma' \frac{m}{2} + \frac{\rho^2}{12} \Sigma' m^2 - \frac{\rho^4}{720} \Sigma' m^4 + \cdots\right) = 0,$$

whence

$$\rho\left(\Sigma n + \Sigma' \frac{m}{2}\right) = \Sigma' 1 + \frac{\rho^2}{12} \Sigma' m^2 - \frac{\rho^4}{720} \Sigma' m^4 + \cdots, \text{ and}$$

$$\rho = \frac{\Sigma' 1 + \frac{\rho^2}{12} \Sigma' m^2 - \frac{\rho^4}{720} \Sigma' m^4 + \cdots}{\Sigma\left(n + \frac{m}{2}\right)},$$

where we have defined $m$ as zero for every lost case. A first approximation to the maximum likelihood estimate is obtained by neglecting all terms after the first in the numerator. This approximation is the number of eventful cases divided by the total number of person-observation time units, using the "$\frac{1}{2}$-rule" for eventful cases.

We can improve the approximation by transferring the term in $\Sigma' m^2$ to the left before division. The result is

$$\rho \doteq \frac{\Sigma' 1}{\Sigma[(n + \frac{m}{2}(1 - \frac{\rho m}{6})]}$$

which neglects terms in $(\rho m)^4$ and higher. If $\rho m < 2$, as will be the case whenever the observations are close enough together for the chance of the event's occurrence between them to be no more than 86 per cent (otherwise we have clearly learned little from the case), then

$$\frac{m}{3} \leq \frac{m}{2}(1 - \frac{\rho m}{6}) \leq \frac{m}{2}$$

and we see that the first approximation was quite good and in usual cases will be entirely satisfactory. Thus, the very simple technique of using $n + \frac{1}{2}m$ as "exposed to risk" will give us a good estimate of $\rho$.

The appearance of $n + \frac{1}{2}m$ has two natural interpretations:

(1) The event "effectively occurs" halfway between last before and first after, thus the case is exposed to risk for half of this period;

(2) "Half of the case" is exposed to risk over the whole period from last before to first after.

Although (1) may seem more natural, we shall find, when we go to a more complex case, that (2) is more accurate and more useful.

11. *Case of stepped* $\rho$

We now go one step toward reality, and let $\rho$ vary with the time. We shall not go all the way, however, but shall assume that it is constant in steps. If there is a real $\rho(t)$, it is surely continuous, but in this case the practical difference between a step-function and a continuous function is small.

We adopt the following notation:

$\rho(t) = \rho_k$ in the $k$-th interval,
$n_k$ = time known before event, in $k$-th interval,
$m_k$ = time of intermediate status in $k$-th interval, and
$\Sigma \rho n$, etc. indicates summation over all $\rho_k n_k$.

Thus, for example, if the intervals are $0 < t \leqq 5$, $5 < t \leqq 10$, $10 < t \leqq 15$, and so on, a case last seen before at time 12, and first seen after at time 21 has

$$\begin{aligned} n_1 &= 5, & m_1 &= 0, \\ n_2 &= 5, & m_2 &= 0, \\ n_3 &= 2, & m_3 &= 3, \\ n_4 &= 0, & m_4 &= 5, \\ n_5 &= 0, & m_5 &= 1, \\ n_6 &= 0, & m_6 &= 0, \end{aligned}$$

and so on.

The contribution of a single case to the likelihood is:

*Lost case*

$$e^{-\Sigma \rho n},$$

*Eventful case*

$$e^{-\Sigma \rho n} - e^{-\Sigma \rho (n+m)} = e^{-\Sigma \rho n}(1 - e^{-\Sigma \rho n}),$$

and thus the derivative of the logarithm of the likelihood with respect to $\rho_k$ is

*Lost case*

$$\Sigma(-n_k),$$

*Eventful case*

$$\Sigma\left(-n_k + \frac{m_k}{e^{\Delta} - 1}\right),$$

where the summations are over cases (and not over $k$), and where

$$\Delta = \rho_1 m_1 + \rho_2 m_2 + \rho_3 m_3 + \cdots.$$

If we introduce the notation

$$f_k = \frac{\rho_k m_k}{\rho_1 m_1 + \rho_2 m_2 + \cdots + \rho_k m_k + \cdots},$$

which is the chance calculated by prorating that the event occurred in the $k$-th interval, we find that the maximum likelihood equation is

$$\rho_k \Sigma n_k = \Sigma' f_k \frac{\Delta}{e^{\Delta} - 1},$$

where as before $\Sigma'$ indicates summation over eventful cases, and expanding $\Delta/(e^{\Delta} - 1)$ as before, we obtain

$$\rho_k \Sigma (n_k + \tfrac{1}{2} m_k) = \Sigma' f_k + \frac{1}{12} \Sigma' f_k \Delta^2 - \cdots$$

$$= \Sigma' f_k + \frac{1}{12} \rho_k{}^2 \Sigma' \frac{m_k{}^2}{f_k} - \cdots$$

$$= \Sigma' f_k + \frac{1}{12} \rho_k \Sigma' m_k \Delta - \cdots .$$

And from this we see that

$$\rho_k = \frac{\Sigma f_k}{\Sigma [n_k + \tfrac{1}{2} m_k (1 - \frac{\Delta}{6})]} ,$$

where all $m_k$ are defined to be zero for lost cases. If our data was at all reasonable, $\Delta/6$ will be a rather small quantity. Thus, we shall not be far from the maximum likelihood solution if we use

$$\rho_k = \frac{\Sigma f_k}{\Sigma (n_k + \tfrac{1}{2} m_k)}$$

to fix the estimates of the $\rho_k$. This result is, unfortunately, not as simple as it seems, since $f_k$ depends on the estimated value of $\rho_k$. However, it is not very hard to use iterative techniques, since the denominators are constant.

Notice that we are clearly now making approximations according to interpretation (2), and are treating a case as half exposed to risk between last seen before and first seen after. As an example of the failure of interpretation (1), let us take our previous example with five-unit intervals, $n = 12$, $n + m = 21$. If we treated it as progressing at $n + \frac{1}{2} m = 16\frac{1}{2}$, the contributions to the various intervals would be 5, 5, 5, $1\frac{1}{2}$, 0, while the actual values of $n_k + \frac{1}{2} m_k$ are 5, 5, $3\frac{1}{2}$, $2\frac{1}{2}$, $\frac{1}{2}$, a quite different sequence. If the $\rho_k$ are about 1 per cent per unit, then for such a case $\Delta$ is about 10 per cent and $1 - (\Delta/6) \sim .98$, so that $n_k + 0.49\ m_k$ is the more accurate quantity. This is very close to the $n_k + 0.50\ m_k$ used in the approximate method.

12. *Use of scores with the information matrix*

While the iterative scheme outlined in the last section has great advantages of simplicity and intuitiveness, it assumes data gathered with gaps of reasonable length. Until there is considerable experience with such analyses, there are bound to be differences of opinion as to whether certain bodies of data are reasonable. And there may even be data worth analyzing which are agreed to be unreasonable. For these

reasons it is worthwhile to record a more complicated process than the simple approximation, one which gives maximum likelihood estimates.

A word about the quality of maximum likelihood estimates is perhaps in order. We know in reasonable generality that:

(1) In very large samples, maximum likelihood estimates are, percentagewise, arbitrarily nearly as accurate as possible;

(2) Other estimates can have this property;

(3) In samples of real sizes the other estimates may, for all we know, be better than the maximum likelihood estimates.

What we know about maximum likelihood estimates is closely related to the basic relations due to Fisher, which refer to expectations, variances, and covariances of the "scores," which can be expressed as derivatives of the logarithm of the likelihood. These derivatives depend on the assumed values of the parameter, and their expectations depend on the true values of the parameter. Fisher's results apply when the assumed and true values are the same, and are

$$E\left\{\frac{\partial \log L}{\partial \rho_k}\right\} = 0,$$

$$\operatorname{cov}\left\{\frac{\partial \log L}{\partial \rho_k}, \frac{\partial \log L}{\partial \rho_j}\right\} = E\left\{-\frac{\partial^2 \log L}{\partial \rho_k \partial \rho_j}\right\}.$$

In our case we have

$$S_k = \frac{\partial \log L}{\partial \rho_k} = \Sigma\left(-n_k + \frac{m_k}{e^{\Delta} - 1}\right)$$

$$= -\Sigma n_k + \Sigma w m_k,$$

$$I_{jk} = \frac{\partial^2 \log L}{\partial \rho_j \partial \rho_k} = -\Sigma \frac{m_j m_k e^{\Delta}}{(e^{\Delta} - 1)^2}$$

$$= -\Sigma m_k m_j v,$$

where the weights $w$ and $v$ are characteristic of the individual case, and depend on the assumed value of the parameters, being given by the simple functions of $\Delta$

$$w = \frac{1}{e^{\Delta} - 1},$$

$$v = \frac{\Delta}{(e^{\Delta} - 1)^2},$$

where

$$\Delta = \rho_1 m_1 + \rho_2 m_2 + \cdots$$

depends on the case and the assumed values of the parameters. If we know, numerically, the scores $(S_1, S_2, \ldots)$ and the information $(I_{11}, I_{12}, I_{22}, \ldots)$ for certain assumed values of the parameters, we may easily either

(1) obtain improved estimates of the parameters, or

(2) test the significance of the indicated deviations from the assumed values.

The first computation proceeds by the natural first approximation, namely that if

$$\rho_k \to \rho'_k = \rho_k + \delta_k ,$$

then

$$S_k \to S'_k \doteq S_k - \sum_j I_{kj}\delta_j,$$

an approximation based on taking expectations in

$$\left[\frac{\partial \log L}{\partial \rho_k}\right]_{\rho+\delta} \doteq \left[\frac{\partial \log L}{\partial \rho_k}\right]_{\rho} + \sum_j \delta_j \frac{\partial^2 \log L}{\partial \rho_k \partial \rho_j}.$$

Thus, the choice of $\delta_j$ which will approximately make $S'_k = 0$ (if $S'_k$ were exactly zero the corresponding values $\rho_k + \delta_k$ would be the maximum likelihood estimates) is the solution of the equation

$$S_k = \sum_j I_{kj}\delta_j$$

which may be solved numerically. We shall write the solution

$$\delta_j = \sum_j I^{jk} S_k,$$

where $I^{jk}$ are the elements of the inverse matrix, which satisfy

$$\sum_k I^{jk} I_{km} = \begin{cases} 1, & j = m, \\ 0, & j \neq m. \end{cases}$$

We immediately find

$$\begin{aligned} \operatorname{cov}(\delta_j, \delta_J) &= \sum_{k,K} I^{jk} I^{JK} \operatorname{cov}(S_k, S_K) \\ &\doteq \sum_{k,K} I^{jk} I^{JK} I_{kK} \\ &= I^{jJ}. \end{aligned}$$

Thus, the values of $I^{jJ}$ allow us both to estimate the $\delta_j$ and to estimate their variances and covariances.

Suppose that, taking a simple example, and starting from $\rho_1 = 0.5$, $\rho_2 = 0.5$, $\rho_3 = 0.5$, we found

$$\begin{aligned} S_1 &= 1.5, \\ S_2 &= -1.0, \\ S_3 &= 0.5, \end{aligned}$$

$$\begin{aligned} I_{11} &= 5.0, \\ I_{21} = I_{12} &= 3.0, \\ I_{22} &= 5.0, \\ I_{13} = I_{31} &= 0.0, \\ I_{23} = I_{32} &= 0.0, \\ I_{33} &= 2.0, \end{aligned}$$

then a simple calculation shows that

$$\begin{aligned} I^{11} &= .3125, \\ I^{12} = I^{21} &= -.1875, \\ I^{22} &= .3125, \\ I^{13} = I^{31} &= .0000, \\ I^{23} = I^{32} &= .0000, \\ I^{33} &= .5000. \end{aligned}$$

The indicated adjustments are

$$\delta_1 = .3125(1.5) - .1875(-1.0) + .0000(0.5) = 0.66,$$

$$\delta_2 = -.1875(1.5) + .3125(-1.0) + .0000(0.5) = -0.59,$$

$$\delta_3 = .0000(1.5) + .0000(-1.0) + .5000(0.5) = 0.25,$$

and the adjusted values are

$$\rho'_1 = \rho_1 + \delta_1 = 0.5 + 0.66 = 1.16,$$

$$\rho'_2 = \rho_2 + \delta_2 = 0.5 - 0.59 = -0.09,$$

$$\rho'_3 = \rho_3 + \delta_3 = 0.5 + 0.25 = 0.75.$$

The adjusted value of $\rho'_2$ is impossible, since negative values are not allowed. The values of $\delta$ should be taken as indicating a repetition of process for a new assumed set of parameters, perhaps $\rho^*_1 = 1.0$, $\rho^*_2 = 0.1$, $\rho^*_3 = 0.75$. By recalculating the contributions $S_j$ and $I_{jk}$ of each case, and repeating the whole process, we can obtain a still better approximation to the maximum likelihood solution.

As an example of the significance testing procedure, let us take the

same example, and consider how we can roughly test the reality of the tendency of the rate in the second period to fall below the rates in the first and third periods. To this end, it is natural to inquire concerning

$$S_1 - 2S_2 + S_3$$

which measures such a tendency. In this example, the observed value is

$$1.5 - 2(-1.0) + 0.5 = 4.0,$$

and the estimated variance is the expectation of

$$(S_1 - 2S_2 + S_3)^2 = S_1^2 - 4S_1S_2 + 4S_2^2 + 2S_1S_3 - 4S_2S_3 + S_3^2,$$

namely

$$\begin{aligned} &I_{11} - 4I_{12} + 4I_{22} + 2I_{13} - 4I_{23} + I_{33} \\ &\quad = 5.0 - 4(3.0) + 4(5.0) + 2(0.0) - 4(0.0) + 2.0 \\ &\quad = 15, \end{aligned}$$

so that the deviation is about 1.03 times its estimated standard deviation and is surely far from significance.

By such procedures of iteration to improve estimates and testing to evaluate trends we can handle even quite unfortunate bodies of observation. In particular, we can take the results of the simpler procedure of estimation and investigate the size and significance of the indicated adjustments.

## REFERENCES

[1] Boag, John W. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Jn. Royal Stat. Soc.*, Series B 11 (1949): 15-44, and 9 pp. of discussion.

[2] Cochrane, A. L., H. W. Campbell and S. C. Stein. The value of Roentgenology in the prognosis of minimal tuberculosis. *Jn. Roent. and Radium Therapy*, 61 (1949): 153-165.

[3] Dorn, Harold F. Methods of analysis for follow-up studies. *Human Biology*, 22 (1950): 238-248.

[4] Greenwood, M. Discussion (of [1] above). *Jn. Royal Stat. Soc.*, Series B 11 (1949): 46-47.

# AN ANALYSIS OF HEART MEASUREMENTS OF GROWING BOYS

BY

C. I. BLISS

*Connecticut Agricultural Experiment Station and Yale University, New Haven, Connecticut.*

AND

M. S. YOUNG

*Department of Pharmacology, Yale University, New Haven, Connecticut.*

IN A recent study by Maresh (2) the size of the heart was measured from periodic chest roentgenograms of 71 boys and 57 girls ranging in age from four to more than 20 years. The object of the study was to relate the size of the heart to the age, height and weight of the growing child. Although other measurements were made from the roentgenograms, the transverse diameter was considered the most reliable. This was plotted against age. Because of the increased rate of growth at puberty, the resulting relations were usually curvilinear and dependent upon the general growth pattern. The study differentiated between individuals only insofar as underweight, medium-weight and overweight children were shown to differ in their heart measurements.

An alternate approach may be based upon that used by the students of allometry or relative growth. It has been found repeatedly (3) that the size of an organ or part can be plotted linearly against the size of the entire organism when both measurements are first converted to logarithms and age is ignored. A modification of this approach has been used by Wetzel (4) in studying the growth of children. In order to separate differences between and within the individuals of the present study, the measurements on each child have been handled as a unit and the logarithm of the heart diameter plotted against the logarithm of body size. This approach has been applied to the original measurements on the 71 boys, for which the authors are indebted to Dr. Maresh.

*The relation of heart diameter to body size.* The size of each boy

was measured in terms of his height and weight each time a roentgenogram was taken. At first glance it appeared that greater precision would follow the use of both measurements than of either alone. This possibility was tested upon the data for 12 boys by computing the partial regression equation relating the transverse diameter of the heart in logarithms ($y$) to the logarithm of height ($x_1$) and the logarithm of weight ($x_2$). A few test diagrams relating $y$ to $x_1$ and $x_2$ showed no appreciable departure from straight lines.

The regression equation $Y = a' + b_1X_1 + b_2X_2$ was then computed by standard methods (1), leading to the partial regression coefficients listed in Table 1 for each boy separately. These accounted for 50 to

TABLE 1

*Relation of logarithm of transverse diameter of heart ($y$) to log-height ($x_1$) and to log-weight ($x_2$) as measured by partial regression coefficients $b_1$ and $b_2$ respectively and computed from N readings upon each boy.*

Each $s^2$ has $N - 3$ degrees of freedom.

| BOY NO. | $N$ | REGRESSION COEFFICIENTS | | % OF VARIATION IN $y$ DUE TO | CORRELATION OF $x_1$ AND $x_2$ | RESIDUAL VARIANCE | $F$ FOR VARIANCE ADDED BY | |
|---|---|---|---|---|---|---|---|---|
| | | $b_1$ | $b_2$ | $b_1$ AND $b_2$ | $r$ | $s^2 \times 10^6$ | $x_1$ | $x_2$ |
| 1 | 31 | — 1.35 | .81 | 80.1 | .990 | 352 | 8.21** | 18.38** |
| 2 | 21 | 2.12 | — .69 | 74.3 | .996 | 334 | 4.62* | 2.26 |
| 5 | 37 | — .28 | .27 | 68.8 | .991 | 127 | 1.18 | 4.88* |
| 6 | 26 | 1.41 | — .23 | 75.2 | .989 | 769 | 3.19 | .41 |
| 8 | 16 | 1.49 | — .21 | 85.0 | .992 | 305 | 2.67 | .39 |
| 17 | 29 | .28 | .14 | 76.3 | .984 | 468 | .37 | 1.08 |
| 20 | 31 | — 2.20 | 1.07 | 79.7 | .992 | 568 | 15.39** | 26.22** |
| 28 | 23 | — 1.26 | .67 | 55.5 | .994 | 288 | 3.75 | 5.97* |
| 34 | 24 | — .43 | .46 | 94.6 | .993 | 201 | 1.05 | 10.54** |
| 45 | 28 | — .16 | .23 | 50.3 | .995 | 419 | .04 | .54 |
| 69 | 25 | — .67 | .51 | 70.9 | .990 | 376 | 1.50 | 5.05* |
| 70 | 16 | — .93 | .41 | 58.5 | .989 | 189 | 1.00 | 2.60 |

* $P < .05$.

** $P < .01$.

95 per cent of the total variation in $y$ as given in the fifth column of Table 1. The larger part of the variation in heart size reflected the growth of the boy. Yet for 11 of the 12 boys one of the two regression coefficients (Table 1) was negative, which suggests that the heart *decreased* in size as the boy's height or weight increased, an obviously absurd result.

The key to the discrepancy proved to be the high correlation between $x_1$ and $x_2$ which frequently exceeded .99 (Table 1). Because of this high correlation a value of $x_1$ largely determined $x_2$ in any given boy. Under these conditions either measure could be substituted for the other in determining the relation between heart diameter and body size. In passing the plane represented by $a'$, $b_1$ and $b_2$ of a partial regression equation through what was essentially a single straight line, the position of the plane depended primarily upon the chance variation in $x_1$ and $x_2$.[1] Hence it was impossible by this means to arrive at a unique solution which had any reasonable physical or biological meaning.

If we conclude that heart size in a given boy could be related to either height or weight but not to both, which of the two alternatives was the more efficient? This has been determined by eliminating first $x_1$ and then $x_2$ from the partial regression equation and computing for each boy how much the residual variation in $y$ would be increased (1). Each increase with one degree of freedom was divided by the residual variance ($s^2$) about the partial regression equation, leading to the values for the variance ratio $F$ shown in the last two columns of Table 1. The variation in $y$ attributable to $x_1$ was reduced significantly in only three of the 12 cases and in two of these was less pronounced than that due to $x_2$. Six of the 12 cases, on the other hand, showed a significant effect of $x_2$ upon $y$. It is evident from this sample of 12 boys that in relating the transverse diameter of the heart to the body size of a single individual, only one measure of size was needed and for this purpose body weight was better than height.

In comparing different boys, however, a much looser relationship would be expected between height and weight as, for example, in boys of the same age. Differences in body build might have made both measurements of value for estimating the "normal" transverse diameter of the heart from a single set of measurements on a given boy. This possibility was tested upon 28 boys from the present series who were measured at both 6 and 12 years of age. Here a much smaller percentage of the variation in $y$ was accounted for by $x_1$ and $x_2$ and the two body measurements showed a much lower correlation, as is evident from the fifth column of Table 2. Nevertheless, the increase in the variation in $y$ due to omitting $x_1$ did not approach significance while that of $x_2$ in one case was significant and in the other exceeded its error. We may conclude

[1] The authors are indebted to Dr. Fredrick Mosteller for pointing out the source of the inconsistency and its geometric and algebraic explanation.

that even in estimating the heart size from a single set of measurements on a given boy, the weight was the more useful observation and no increase in precision followed the measurement of height as well.

*Relation of heart diameter to body weight.* A straight line was then computed relating the logarithm of the transverse diameter of the heart ($y$) to the logarithm of body weight ($x$) for each of 54 boys. These individuals were represented by records which covered a span of more than 5½ years and continued beyond 9½ years of age or started before 11 years of age. Two boys were omitted which met these conditions (Nos. 66 and 67); they were the only ones in the entire series whose hearts did not show a significant increase in size over an observational period lasting more than six years. On the average the boys meeting these requirements were measured 30 times, the number varying from 14 to 43 per boy.

TABLE 2

*Relation of heart diameter to height and weight in same units as in Table 1, when computed from the reading on each of 28 boys at age 6 and again at age 12.*

| AGE IN YEARS | REGRESSION COEFFICIENTS | | % OF VARIATION IN $y$ DUE TO | CORRELATION OF $x_1$ AND $x_2$ | RESIDUAL VARIANCE | $F$ FOR VARIANCE ADDED BY | |
|---|---|---|---|---|---|---|---|
| | $b_1$ | $b_2$ | $b_1$ AND $b_2$ | $r$ | $s^2 \times 10^6$ | $x_1$ | $x_2$ |
| 6 | —.054 | .272 | 19.4 | .770 | 612 | .01 | 2.69 |
| 12 | —.290 | .353 | 28.4 | .564 | 871 | .33 | 8.30** |

** $P < .01$.

To illustrate the relation of $y$ to $x$ the results on six representative boys have been plotted in Fig. 1. In the cases which were examined in this way there was no consistent departure from a straight line which would warrant fitting some other type of curve. Moreover, the variability in heart diameter was independent of body weight when both were expressed in logarithms. In consequence a straight line was computed by least squares from the data for each boy in the form $Y = a + b(X - \bar{x})$, where $Y$ and $X$ are predicted or assigned values, $b$ is the slope of the straight line and $a = \bar{y}$ or the mean of $y$ at $\bar{x}$, the mean of $x$. The variance about each line ($s^2$) was determined by the usual method (1).

The positions of the lines were compared at a point where they would be little affected by the error in slope, at a value of $X$ which approached the unweighted mean of the $\bar{x}$'s for the 54 straight lines in the series. For this reason the expected log diameter of the heart ($Y$)

was computed from the straight line for each boy at $X = 1.4771$ or the logarithm of 30 kg. The constants computed from the data for each boy are exemplified in Table 3 for the six boys plotted in Fig. 1. From the

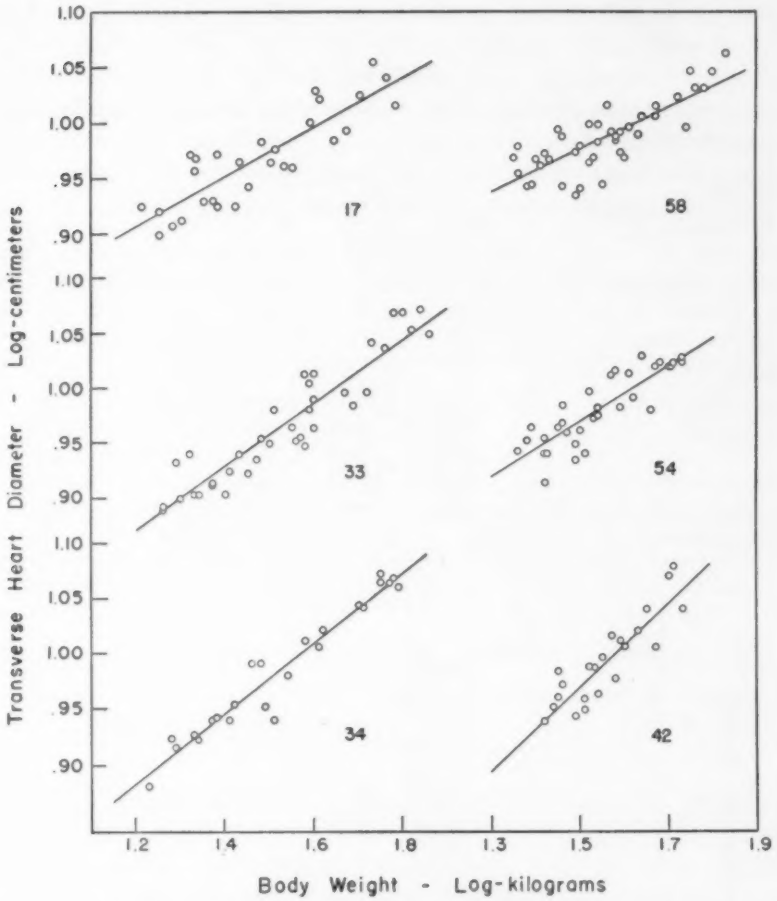


FIG. 1. RELATION OF TRANSVERSE DIAMETER OF THE HEART TO BODY WEIGHT IN SIX GROWING BOYS FROM MARESH'S STUDY (2), BOTH COORDINATES IN LOGARITHMIC UNITS.

The statistics of these curves are given in Table 3.

variability in constants such as these the reliability of the measurements of heart size has been examined.

The heart measurements varied considerably about the growth curve of

any individual as may be observed in Fig. 1. The standard deviation ($s$) for the residual variation about the line fitted to the data for each boy has been plotted as a frequency histogram in Fig. 2. In terms of $s$ this varied from 0.011 to .031, a greater scatter about the combined value ($s_c = .0205$) than would be expected by chance ($\chi^2 = 124.4$,

TABLE 3

*Statistics for relation of transverse diameter of heart to body weight for observations on six boys plotted in Fig. 1.*

| BOY NO. | 17 | 33 | 34 | 42 | 54 | 58 |
|---|---|---|---|---|---|---|
| No. of observations, $N$ | 29 | 37 | 24 | 22 | 33 | 41 |
| Age span, years | 4.0-13.5 | 4.8-16.0 | 5.9-19.0 | 8.3-14.0 | 8.3-17.0 | 5.5-18.5 |
| Geometric mean: Weight, kg. | 29.7 | 34.7 | 33.5 | 36.2 | 35.0 | 35.8 |
| Geometric mean: Heart, cm. | 9.27 | 9.30 | 9.67 | 9.83 | 9.57 | 9.67 |
| Size of heart at 30 kg. | 9.29 | 8.92 | 9.34 | 9.15 | 9.20 | 9.35 |
| Logarithms: | | | | | | |
| Mean weight, $\bar{x}$ | 1.4724 | 1.5400 | 1.5246 | 1.5591 | 1.5442 | 1.5541 |
| Mean heart, $\bar{y}$ | .9669 | .9686 | .9854 | .9926 | .9807 | .9856 |
| Slope, $b$ | .2249 | .2868 | .3170 | .3805 | .2533 | .1941 |
| Standard deviation, $s$ | .0214 | .0216 | .0142 | .0187 | .0172 | .0185 |
| Expected $Y$ at $X = 1.4771$ | .9680 | .9506 | .9703 | .9614 | .9637 | .9707 |
| Residual variation, % | 24.1 | 15.2 | 5.7 | 21.2 | 26.3 | 34.3 |

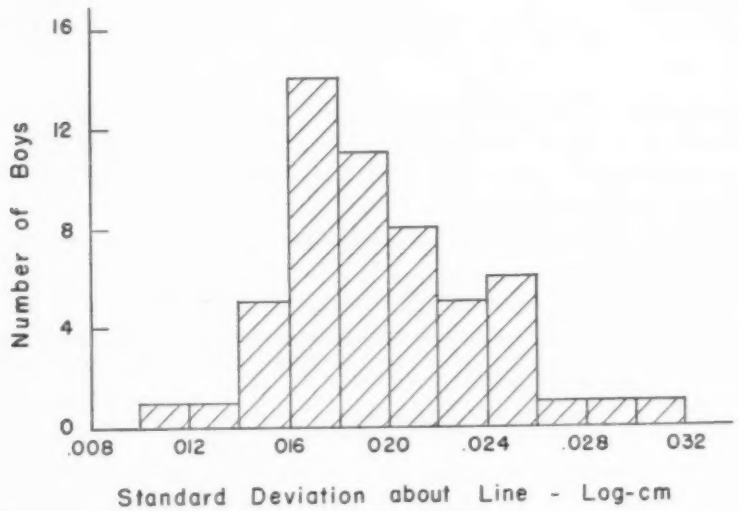


FIG. 2. DISTRIBUTION OF THE STANDARD DEVIATION IN $y$ ABOUT THE FITTED STRAIGHT LINES FOR 54 BOYS

$n = 53$). The residual variation about the fitted lines varied in the different boys from 6 to 58 per cent of the total variation (sum of squares) in $y$, with a median value of 29 per cent. The author (2) comments on the fluctuation in heart measurements from film to film and suggests that some of it "may be due to the different phases of the cardiac cycle during which the roentgenograms were taken." The measured diameter of the heart may also vary with changes in position, degree of tension, depth of inspiration, intrathoracic pressure and other physiological factors. These might conceal or distort the course of normal growth in a given boy at any particular time. Only the line computed from many roentgenograms would average out such differences.

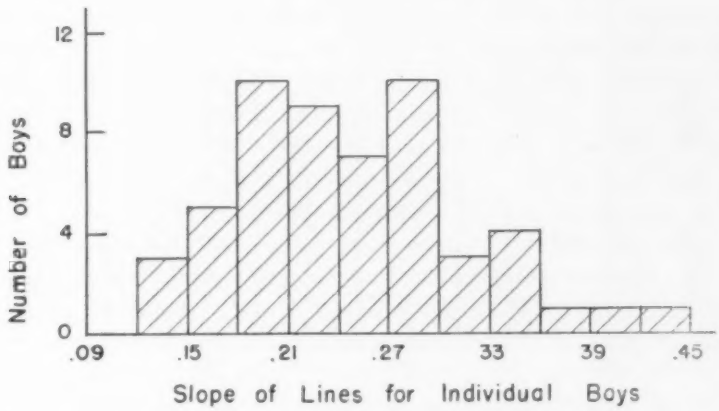


FIG. 3. DISTRIBUTION OF THE SLOPES ($b$) OF THE STRAIGHT LINES FOR 54 INDIVIDUAL BOYS

In studying the relation of the growth of a part to the growth of the whole the statistic of greatest interest is the slope of the line relating their logarithms. If the diameter of the heart were to increase at the same rate as the weight of the body, this slope would approach one-third, on the assumption that weight was proportional to the third power of a linear measurement. In the present observations, however, it ranged from 0.140 to 0.431 with a weighted average of $b_c = 0.242$, which indicates a smaller rate of increase in the heart than in the body as a whole. The variation in this rate of increase among the 54 boys in the series is shown in the frequency histogram in Fig. 3.

Analysis of variance revealed a greater variation in slope from boy

to boy than would be expected from the average variability of the observations around the fitted lines ($P < .001$). The discrepancy is also shown by the greater observed standard deviation among the unweighted slopes ($s_b = .0670$) than the square root of the average variance for the individual slopes ($\bar{s}_b = .0311$). Hence in this study the rate of growth of the heart relative to that of the body varied significantly from one individual to another.

The individual readings for the six boys in Fig. 1, and for the other series which were plotted similarly, seemed on inspection to be fitted satisfactorily by straight lines. If the relation were linear in fact, the relative growth rate of the heart should be independent of the weight

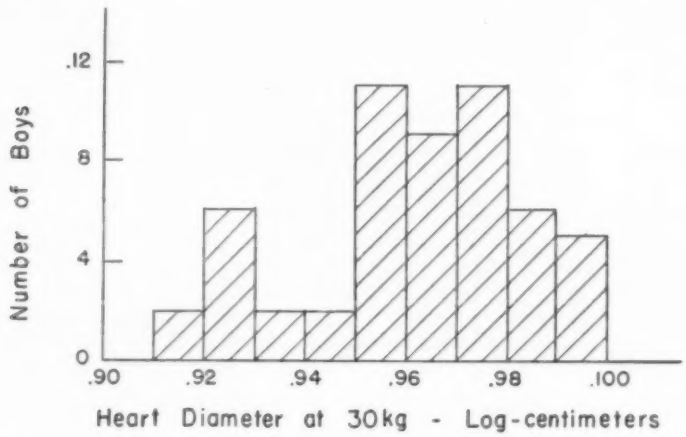


FIG. 4. DISTRIBUTION OF THE LOG-HEART DIAMETER ($Y$) AS COMPUTED FROM INDIVIDUAL CURVES AT A BODY WEIGHT OF 30 KG.

(or age) of the boy and the slope ($b$) of the computed lines should not vary with the mean log-weight ($\bar{x}$). However, when this was tested graphically by plotting $b$ against $\bar{x}$ for each of the 54 boys, $b$ tended to increase with $\bar{x}$. This trend proved to be statistically significant ($P < .001$) and accounted for 26 per cent of the observed variation among the unweighted slopes. The present data were not suitable for determining whether the increase in the relative growth rate of the heart was a continuous or a discontinuous function of weight (or age).

The diameter of each boy's heart when he weighed 30 kg. was computed in logarithms from the line based on his individual readings. These values ($Y$) ranged from 0.1102 to 0.1998 with the distribution shown by the frequency histogram in Fig. 4. For the same body weight, therefore,

the diameter of the largest heart was only 23 per cent greater than that of the smallest heart in the series. The fastest relative rate of heart growth, on the other hand, was 95 per cent greater than in the heart which grew the slowest in this series of 54 boys, as measured by their respective slopes. This suggests superficially that the individual curves differed more in slope ($b$) than in position ($Y$).

In both cases, however, the critical comparison is that of the observed with the expected standard deviation of the statistic and in terms of this criterion the heart diameter at 30 kg. ($Y$) was twice as variable as its relative rate of growth ($b$). The observed unweighted standard deviation in $Y$, as computed from the frequency distribution in Fig. 4, was $s_Y = 0.02305$. Its unweighted mean standard error, computed (1) as the square root of the mean variance of $Y$ for $X = 1.4771$, was $\bar{s}_Y = 0.00509$. Hence the values of $Y$ had an observed standard deviation that was $.02305/.00509 = 4.53$ times as great as would be predicted from the inherent precision of the individual curves. The corresponding ratio for the slopes was $.0670/.0311 = 2.15$ or less than half that observed for the positions of the same lines. We may conclude, therefore, that the boys' hearts were in closer agreement in terms of relative growth rate than of diameter at a given body weight. Unlike the slopes, the values of $Y$ seemed to be independent of the mean log-weight $\bar{x}$.

Finally, the mean values of $Y$ and $b$ may be combined into a single general equation for estimating heart diameter from body weight in growing boys. Because of the indirect evidence that the relation in log units was not really linear, a general equation based upon the present data is of limited reliability. So far as could be judged from the means of the statistics for 54 separate curves, the general equation may be written as

$$Y = 0.9617\ [\pm .0230] + 0.2478\ [\pm .0670]\ (X - 1.4771),$$

where $Y$ is the log-transverse diameter of the heart (in cm.) averaged over the cardiac cycle, $X$ is the log body weight (in kg.) and [ ] enclose the standard deviations of the unweighted statistics from 54 curves.

This equation may be applied to the data from two boys whose records have been omitted because they covered a span of only 3.5 years. The equation relating nine measurements of heart diameter to body weight, both in logarithms, was $Y = 0.953 + 0.301\ (X - 1.613)$ for boy No. 32, whose record started at 11.2 years of age. For eight measurements on boy No. 62, $Y = 0.914 + 0.258\ (X - 1.319)$, his record beginning when he was 4.0 years old. In both boys the heart was growing at a normal rate relative to their bodies, their individual slopes falling within one

standard deviation of the combined slope for the main series, 0.248 $\pm$ .067 = 0.181 and 0.315.

Their expected heart diameters at a weight of 30 kg. may be compared similarly to $Y = 0.962 \pm 0.023$. Boy No. 32 weighed 30.2 kg. at the start of his record and boy No. 62 weighed 24.7 kg. when his record terminated. As extrapolations, the log heart diameters expected at 30 kg. were subject to relatively larger errors. When their equations were solved for $X = 1.477$, $Y = 0.912 \pm 0.016$ for boy No. 32 and $Y = 0.955 \pm 0.016$ for boy No. 62. The latter heart measurement agreed well with the main series, but boy No. 32, seemed to have a relatively small heart for his weight, as can be checked by comparison with the frequency distribution in Fig. 4. The difference in $Y$ ($0.050 \pm 0.028$) was not sufficient, however, to consider him inconsistent with the main series.

*Summary.* Measurements of the transverse diameter of the heart from periodic roentgenograms of growing boys have been examined statistically in relation to measurements of body size. In these data heart diameter could be predicted more efficiently from body weight than from either height or a combination of weight and height. The logarithm of the heart diameter ($y$) plotted against the logarithm of body weight ($x$) was fitted by a straight line for each boy. The standard deviation of the individual observations about each line was significantly larger in some boys than in others. The slope ($b$), measuring the relative growth rate of the heart, varied more from boy to boy than would be expected from the standard deviations about the individual lines. An even greater excess variability was observed in the heart diameter ($Y$) computed from each curve for a constant weight (30 kg.). Some of the variation in $b$ but not in $Y$ was associated with the mean weight ($\bar{x}$), suggesting that the slope may prove to be somewhat steeper in larger than in smaller boys.

## REFERENCES

1. Fisher, R. A. Statistical Methods for Research Workers. 10th ed., Hafner Publishing Co., Inc. (*New York*), 1948.
2. Maresh, M. M. *Pediatrics*, 2: 382-404. 1948.
3. Teissier, G. *Biometrics*, 4: 14-48. 1948.
4. Wetzel, N. C. Medical Physics, pp. 530-568. Year Book Publishers, Inc. (*Chicago*), 1944.

# MARRIAGE FERTILITY BY AGES OF COUPLES

BY THOMAS P. MONAHAN

PRINCIPALLY interested in the procreation of the largest number of healthy children, the philosophers Plato and Aristotle defined the limits and the desirable combinations of age for parenthood in their ideal states.[1] There has developed in the last century a new interest in the subject of the influence of age and joint ages upon the fertility, fecundity, and sterility of couples; the quality of the offspring, its health, intelligence, and longevity; and the sex ratio and plurality of births. Age disparity of couples has also been related to the likelihood of widowhood and orphanhood, to sexual adjustment, and to the happiness and probable success of marriage. And, in the last few years, essential consideration is given to the marital condition of the population and age disparity in the determination of the true rate of population increase.[2]

One of the first major investigations of the subject of fertility according to age was that by J. M. Duncan in 1866,[3] and since then many other statistical investigations have been made. It should be

[1] Aristotle. *Politics*, trans. by Jowett. *New York*, Random House, The Modern Library, 1943, 313-16.
Plato. *The Republic*. *Ibid.*, 181, 184.
———. *The Laws*. *Ibid.*, cix, 104.
Roper, A. G. *Ancient Eugenics*. *Oxford*, Blackwell, 1913, 55, 61-62.
Savage, C. A. *The Athenian Family*. *Baltimore*, Lord Baltimore Press, 1907, 50-51.
Crum, F. S. The statistical work of Süssmilch. *Q. Pub. Amer. Statis. Assn.*, 7: 17, 1901.

[2] Karmel, P. An analysis . . . net reproduction rates in actual populations. *Population Studies*, 2: 246, 256-57, 273, 363, 1948-49.
Hajnal, J. Some comments on Mr. Karmel's paper. *Ibid.*, 357-58.
Stolnitz, G. and N. Ryder. Recent discussion of the net reproduction rate. *Population Index*, 114-28, Apr. 1949.

[3] Duncan, J. M. Fecundity, Fertility, Sterility, and Allied Topics. *Edinburgh*, A. &. C. Black, 1866.

recognized at the outset that none of the many studies measures change in biological capacity alone. Physiological impairment, artificial birth control and continence, and social factors also combine to influence any observable change in fertility with age. On the basis of some highly refined data Pearl has shown how slightly the biological function alone may actually decline with advancing age.[4]

The publication of a study of data for Budapest on fertility and joint ages by Körösi in 1894 renewed interest in this aspect,[5] although Sadler and others had presented smaller studies over 50 years earlier.[6] Other investigators pursued the subject further, and in 1911 a number of national censuses assembled population data on fertility according to duration of marriage and relative ages.[7] In this country no study comparable to Körösi's has been made. The few studies that have been made are highly limited, dealing with the question only briefly or using

[4] Pearl, Raymond. The Natural History of Population. *New York*, Oxford Univ. Press, 1939, 36-42, 156-58.

[5] Körösi, J. An estimate of the degree of legitimate natality, etc. *Proc. Roy. Soc. of London*, 55: 16-23, 1894. Also published in (1) *J. Roy. Statis. Soc.*, 57: 690-708, 1894, and (2) *Philos. Trans. Roy. Soc. of London*, Vol. B, Pt. 2: 781-875, 1895.

See also the following: W. Bailey, Modern Social Conditions, *New York*, Century Co., 1906, 185, 190 and others; A. Newsholme, The Elements of Vital Statistics, *New York*, Appleton, 1899, 67; and G. Whipple, Vital Statistics, 2nd ed. *New York*, Wiley, 1923, 249.

[6] Quetelet, M. A. A Treatise on Man, transl. *Edinburgh*, Chambers, 1842, 13 and others. See also: Körösi, *op. cit.*

[7] *Population Index* (Quarterly). School of Public Affairs, Princeton Univ.

Coghlan, T. A. Childbirth in New South Wales. *Sydney*, Govt. Pr., 1899.

See also: *J. Roy. Statis. Soc.*, 157-59, 1899; 133-35, 1906.

Lewis, C. J. and J. N. Lewis. Natality and Fecundity. *London*, Oliver & Boyd, 1906.

Ewart, R. J. The aristocracy of infancy and the conditions of its birth. *Eugenics Rev.*, 3: 143-69, July 1911.

———. The influence of parental age on offspring. *Eugenics Rev.*, 3: 201-32, Oct. 1911.

Dunlop, J. C. The fertility of marriage in Scotland, a census study. *J. Roy. Statis. Soc.*, 77: 259-88, 1913-14.

England, Census of 1911. Fertility of Marriage, Pt. 1.

Scotland, Census of 1911. Vol. III: xxxviii.

Muller, J. Human fertility in relation to ages of husband and wife at marriage and duration of marriage. *An. Eugenics*, 4: 238-78, 1931.

Australia, Census of 1911. Vol. 1, 262 and others; and Appendix A, 1-466.

highly specialized samples.[8] Following the European example, the United States in 1940 published a series of reports on differential fertility which revealed unilinear aspects of fertility with age, but which ignored the question of relative ages.[9] Ever since 1917 the United States has published birth data according to relative ages of parents,[10] but very little use has been made of these materials because the necessary population data were never compiled and rates therefore could not be calculated.

## DATA

For the United States the question of age and age disparity as related to fertility of couples may be brought partially into focus by some data that have been published in recent years but not heretofore analyzed.[11] More detailed data could be laboriously compiled from old census and birth records, but it is hoped that the results of this paper will encourage the Federal agencies to assemble additional and more adequate tables from current data. Oddly enough, now, just when the United States Census Bureau is beginning to publish population data in this manner the birth tabulations have been considerably abbreviated.

[8] Pomerat, G. R. Fertility in relation to age at time of marriage. *Human Biol.*, 8: 426, 431, 1936.
McKain, W. C. and N. Whetten. Size of family in relation to homogeneity of parental traits. *Rural Soc.*, 1: 20-27, Mar. 1936.

[9] U. S. Bureau of the Census. Population, Differential Fertility, 1940 and 1910. 5 vols.
Notestein, F. W. Intrinsic factors in population growth. *Proc. Amer. Philos. Soc.*, 80: 499-511, 1939.
Kiser, C. V. and P. K. Whelpton. Social and psychological factors affecting fertility. *Milbank Memorial Fund Quarterly*, 21: 232-33, 1943. See also subs. vols.

[10] U. S. Bureau of the Census. Birth Statistics, 1915, and others.

[11] Data for each year obtained from reports instanced:
1940 (a) U. S. Census of 1940, Population, Vol. IV, 25.
(b) Current Population Report, P20, No. 10, 10-11.
(c) Types of Families, 1940, Spec. Rpt., 3.
(d) Vital Statistics of the United States, Pt. 2, 110.
1946 (a) Current Population Report, P20, No. 8, 10-11.
(b) Vital Statistics of the United States, Pt. 2, 118.
1947 (a) Current Population Report, P20, No. 10, 10-11.
(b) Vital Statistics of the United States, Pt. 2, 100, 165.
1948 (a) Current Population Report, P20, No. 26, 18.
(b) Advance report on the 1948 birth data from Mr. Shapiro.

In 1940 and 1948 the United States Census reported on the relative ages of married couples.[12] Births according to relative ages of parents appear in the published report on *Vital Statistics of the United States*, and an advance copy of the same tabulation for 1948 was obtained from Mr. Sam Shapiro of the National Office of Vital Statistics. To obtain the fertility rate by relative ages of married couples, all that was necessary was to divide the births in each age-cell by the number of husband and wife families approximately corresponding in 1940 and 1948.

No claim is made here for the precision of the rates; on the contrary their approximate nature should be fully recognized. Persons familiar with such statistical compilations are most aware of the deficiencies, especially regarding the accuracy of reporting. The unilinear age rates are based upon all husband-wife families, representing a complete count in 1940 and an estimate of their number for 1946 to 1948.[13] The sample surveys were made in June, April and April of the respective years, covering about 25,000 households.[14] The tabulation of relative ages in 1940 was from a large sample ($2\frac{1}{2}$ to 5 per cent) of the total population, but rather oddly represents only about 93 per cent of all husband-wife families,[15] thus magnifying the calculated fertility rates by relative age in 1940, and making them non-comparable with those for 1948 which include all husband-wife families in the sample. Because of the size of the population sample in 1948 the relative age rates are not to be trusted in all age-cells.

In the birth tabulations illegitimate births were not separated, while the population base used to calculate the rates included husband-wife families only, so that the rates are higher than truth. Interpretation of the rates must allow for distortion due to the inclusion of illegitimate births, a thing which Körösi had apparently not done in his Budapest study.[16] The relative age pattern of illegitimacy may really be more dispersed than for customary unions,[17] and, if this is so, one could expect their intermingling in the birth (but not in the population) data to mask rather than exaggerate differences in the relative rates.

[12] Reference 1940c and 1948a, above.

[13] 1940 population from reference 1940a and 1940b, above.

[14] For an appraisal of sampling errors and the like, see Current Population Reports.

[15] Reference 1940c, 3, above.

[16] Coghlan, *op. cit.*, 34. Cf.: A. O. Powys, Data for the problem of evolution in man, *Biometrika*, 4: 259, 263, and 1: 33.

[17] Philadelphia Municipal Court, Annual Report. Occasional tables on ages of illegitimacy and bastardy cases.

Furthermore, the birth data include a sizable number of fathers, and a small number of mothers, whose ages were not given, with no indication of the proportion of illegitimates therein. About 3 per cent of all fathers did not give their age; but, for mothers under 20, the fathers' ages were not reported in about 10 per cent of the cases. These unknown ages were eliminated by prorating them into the table of births, adjusting the fathers' ages first in the relative age tables. This is a reasonable but mathematically arbitrary device. Husbands of all ages are represented, but wives up to age 54 only, assuming that no births occurred to women after that age. Unfortunately, except for the unilinear rates in 1947, the data are by 5-year age groups only, with the relative age data even more telescoped in 1940. If population and birth data were ever assembled to show relative ages by single years, and further refined for illegitimacy, first marriage, order of birth and so on, along with the other customary divisions of color and locality, some very significant results could be obtained which would disclose more clearly than is now possible the pattern of fertility by age in the United States.

## ANALYSIS OF UNILINEAR AGE DATA

The fertility rates of married couples show an uninterrupted decline from the earliest age group onward, for males and females alike. If husbands over 64 years were excluded the female rates would be raised only slightly. The picture differs drastically from that conveyed by published fertility rates in general use, which are based upon all women.[18] How the two kinds of rates differ is shown in Table 1.

The rise in fertility up to age group 20-24, based upon all women, is a fictitious one. Since a large number of women in the first age group are unmarried and not participating in the birth pattern they should not be included in the calculation of a rate. The fertility rates based upon husband-wife families are much higher at the start, and the gradient to the end of the reproductive period much steeper therefore. In corresponding age groups, except for the under 20 group, the unilinear marital fertility rates of the males exceed those for the females, but most of this can be explained as being due to the customary age-difference in marriage, which approaches 4 years. By lagging the male rates one age group behind the female rates, we find that the female rate is higher until the age of wife of 35 years is passed, whereafter the rate for the

[18] Vital Statistics Special Reports. Vol. 35, No. 10, 131.

husband is higher. At age 45-49 the wife's fertility has almost ceased, but about 7 in 1,000 of the husbands 55 years and over become fathers each year. It is remarkable that although roughly 40 to 50 per cent of the married women have a child within the year, the decline with each succeeding age group is almost a geometric ratio of .6, or a rate of decline of 40 per cent. The gradient for the males is somewhat less steep and the curve asymptotic, no doubt.

Allowing for illegitimacy mostly alters the first part of the female fertility curve, since the proportion of births which are illegitimate

TABLE 1

*Fertility rates of females by age*[19]

Live births per 100 women and per 100 married women (husband present). Population base: 1940, census; 1948, estimate from sample.

| AGE | 1940 All women | 1940 Married women | 1948 All women | 1948 Married women |
|---|---|---|---|---|
| Under 20 | 4.89 * | 46.45 | 7.97 * | 53.77 |
| 20-24 | 12.50 | 25.95 | 19.28 | 31.30 |
| 25-29 | 11.41 | 16.27 | 16.03 | 20.08 |
| 30-34 | 7.71 | 10.14 | 10.03 | 12.41 |
| 35-39 | 4.18 | 5.43 | 5.33 | 6.62 |
| 40-44 | 1.39 | 1.82 | 1.51 | 1.97 |
| 45-49 | 0.13 | 0.18 | 0.11 | 0.15 |
| 50-54 | 0.00 | 0.01 | 0.00 | 0.01 |
| 10-54 | 5.19 | 9.85 | 7.42 | 12.38 |
| 15-44 | 7.35 | | 10.48 | |

* Based on women 15-19.

decreases rapidly with age of mother. In 1940, it was estimated, about two-thirds of the births to mother aged 10-14, and one-eighth of those aged 15-19 were illegitimate.[20] The published estimates of illegitimate births by age enables us to derive adjusted female marital fertility rates for 1940, 1946 and 1947, shown in Table 2.

Fertility rates based upon all women (or all men) can produce another very misleading result, for a change in such rates from one time to another may merely mean a change in the proportion married (as during a depression, or during and following a war) and not any

[19] Footnote 18 and reference 1940a, 1940b, 1940d, and 1948a and 1948b, above.
[20] Vital Statistics Special Reports. Vol. 33, No. 5, 73-74.

TABLE 2

*Marital fertility rates*[21] *adjusted for illegitimate births*

Births, excluding stillbirths, per 100 married women (husband present). Complete population count, 1940; sample, 1946 and 1947. The first age group in 1946 is for 15-19 years.

| | 1940 | | 1946 | | 1947 | |
|---|---|---|---|---|---|---|
| | Unadj. | Adj. | Unadj. | Adj. | Unadj. | Adj. |
| All ages to 54 | 9.85 | 9.48 | — | — | 13.36 | 12.88 |
| Under 20 | 46.45 | 39.96 | 60.31 | 51.17 | 61.88 | 53.88 |
| 14-17 | | | | | 76.10 | 58.22 |
| 18-19 | | | | | 57.46 | 52.54 |
| 20-24 | 25.95 | 25.00 | 34.10 | 32.71 | 34.64 | 33.41 |
| 25-29 | 16.27 | 16.00 | 22.08 | 21.68 | 22.63 | 22.24 |
| 30-34 | 10.14 | 10.00 | 13.29 | 13.11 | 13.60 | 13.41 |
| 35-39 | 5.43 | 5.34 | 7.17 | 7.06 | 7.15 | 7.04 |
| 40-44 | 1.82 | 1.80 | 2.00 | 1.97 | 2.02 | 1.98 |
| 45-54 | .10 | .10 | .10 | .09 | .09 | .09 |

TABLE 3

*Index of change of female fertility rates from 1940*

Including illegitimates, excluding stillbirths.

| AGE GROUP | MARITAL FERTILITY | | | ALL WOMEN |
|---|---|---|---|---|
| | 1946 | 1947 | 1948 | 1948 |
| 14-54 | — | 1.36 | 1.26 | 1.43 |
| Under 20 | 1.30 | 1.33 | 1.16 | 1.63 |
| 20-24 | 1.31 | 1.33 | 1.21 | 1.54 |
| 25-29 | 1.36 | 1.39 | 1.23 | 1.40 |
| 30-34 | 1.31 | 1.34 | 1.22 | 1.30 |
| 35-39 | 1.32 | 1.32 | 1.22 | 1.28 |
| 40-44 | 1.10 | 1.11 | 1.08 | 1.09 |

[21] *Ibid.*, and footnote 11 (except 1940d).

real change in the fertility pattern.[22] The change in the unilinear age-specific marital fertility rates was proportionately evidenced in all age periods of females, as shown in the following summary (Table 3), and not concentrated in any one age group of females. It is unlikely, therefore, that this change at each age group has been due to an increase in first births alone.

The age-specific marital fertility rates were higher for women (and men) in the three years 1946-48 than in 1940, but they had not increased nearly as much as the age-specific rates for all women would make it appear. In the years 1940 and 1946-48, according to the age-specific marital fertility rates, counting the age period under 20 as of 3 years duration only, 100 women passing through the reproductive period (none dying) would leave behind them 438, 575, 587, and 522 children, respectively. The ratio of change in these hypothetical offspring might be called a gross marital reproduction index.[23]

TABLE 4

*Indexes of change in fertility since 1940* [24]

| | 1940 | 1946 | 1947 | 1948 |
|---|---|---|---|---|
| Gross reproduction rate | 1.12 | 1.44 | 1.61 | 1.54 |
| Net reproduction rate | 1.02 | 1.36 | 1.52 | 1.46 |
| Gross marital reproduction | 438 | 575 | 587 | 522 |
| *Ratio to 1940* | | | | |
| Gross reproduction rate | | 1.29 | 1.44 | 1.37 |
| Net reproduction rate | | 1.33 | 1.49 | 1.43 |
| Gross marital reproduction index | | 1.31 | 1.34 | 1.19 |

The standard gross and net reproduction rates (derived from age specific rates of all women) in 1947 and 1948 do not take into account the large increase in marriages during 1946 and 1947; [25] hence these

[22] Whelpton, P. K. The meaning of the 1947 baby boom. *Vital Statistics Special Reports*, Vol. 33, No. 1.

———. Is family size increasing? *Vital Statistics Special Reports*, Vol. 23, No. 16.

Karmel, P. H. A note on P. K. Whelpton's calculation of parity adjusted reproduction rates. *J. Amer. Statis. Assn.*, 45: 119-24, 1950.

Whelpton, P. K. Comments on Mr. Karmel's note. *Ibid.*, 125-35.

[23] A different approach to the concept of marital fertility is suggested in Current Population Reports, P20, No. 8, and P20, No. 27.

[24] See Vital Statistics Special Reports, Vol. 35, No. 10, 133.

[25] *Ibid.*, Vol. 31, No. 16.

rates increased and remained discernibly higher than in 1946. As revealed by the gross marital reproduction index, there was, on the contrary, no significant change in marital fertility from 1946 to 1947, and in 1948 there was probably an important decline in the fertility of married couples. If these data are reliable, and if the question of first births is also an important further qualification, then there may have been a real decline in the fertility of marriage since 1946.

### ANALYSIS OF RELATIVE AGE DATA

Turning to the rates by relative ages (Tables 5 and 6) we find some new information suggested by the calculations. As already indicated, the 1940 rates are higher because of the underenumeration of the population base of husband-wife families by relative age, and are not to be compared to the rates for 1948.

In each of the age-lines of husband and wife, from the earliest age group onwards, the fertility rate declines regularly with advancing age, indicating fertility is a function of the husband's age as well as the wife's age. It appears, as Galton has written, that fertility is determined by joint age,[26] and not by the age of the mother alone. In a particular age group, men having older wives show increasingly lower fertility rates, and the same result obtains for women having increasingly older husbands. On the other hand, when men and women of a specified age group have increasingly younger spouses the fertility rate rises; from the wife's point of view, her fertility rises the more she is older than her husband; from the husband's point of view, his fertility rises the more he is older than his wife. For the wife, an older husband means lower fertility; and for the husband, an older wife means lower fertility also. The loss of fertility with increasing age can only be partly recovered by enlarging the age-difference of the pair; and for persons who start procreating later in life it is a stern fact that the span of fertile years ahead of them is shorter than for those who start early. Late starters have a lower productivity of births and a smaller span of years for birth spacing.[27]

A few departures from the general bilateral pattern can be observed. In the results for 1940, for wives aged 35 years and over, the fertility rate increases from the under 25 to the 25-29 age group of husbands,

[26] Galton's comment in Körösi, *op. cit.*, *J. Roy. Statis. Soc.*, 705, 1894.

[27] See Notestein, footnote 9 above.

and indeed remains higher for husbands in the next higher age group who have wives aged 40-44. And for husbands with wives aged 20-24 the rate does not decline as it might be expected from age group 30-34

TABLE 5

*Marital fertility rates by relative age,[28] United States, 1940*

Births, excluding stillbirths but including illegitimates, per 100 married couples; wives under 55 years.

| AGE OF HUSBAND | AGE OF WIFE | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Total | Under 20 | 20-24 | 25-29 | 30-34 | 35-39 | 40-44 | 45-49 | 50-54 |
| Total | 10.64 | 65.28 | 31.49 | 18.05 | 10.87 | 5.67 | 1.89 | .18 | .01 |
| Under 25 | 46.30 | 72.48 | 38.59 | 27.56 | 20.86 | 11.99 | 2.90 | .74 | .14 |
| 25-29 | 25.12 | 54.78 | 29.48 | 20.38 | 16.72 | 12.36 | 5.72 | 1.97 | .92 |
| 30-34 | 15.64 | 49.37 | 26.93 | 16.70 | 12.14 | 8.98 | 3.85 | .88 | .44 |
| 35-44 | 7.21 | 42.88 | 26.38 | 15.36 | 9.62 | 5.58 | 2.15 | .30 | .10 |
| 45-54 | 1.76 | 27.74 | 20.57 | 14.31 | 8.78 | 4.88 | 1.67 | .18 | .01 |
| 55 and over | .72 | 13.25 | 13.21 | 11.70 | 8.38 | 5.12 | 1.56 | .13 | .00 |

TABLE 6

*Marital fertility rates by relative age,[29] United States, 1948*

Births, excluding stillbirths but including illegitimates, per 100 married couples; wives under 55 years. Rates based on an estimated population of 5,000 or less not shown; ( ) indicates a population base of less than 10,000.

| AGE OF HUSBAND | AGE OF WIFE | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Total | Under 20 | 20-24 | 25-29 | 30-34 | 35-39 | 40-44 | 45-49 | 50-54 |
| Total | 12.38 | 53.77 | 31.30 | 20.08 | 12.41 | 6.62 | 1.97 | .15 | .01 |
| Under 20 | 50.78 | 50.02 | 52.51 | | | | | | |
| 20-24 | 38.83 | 56.40 | 34.96 | 24.99 | 26.49 | (12.94) | | | |
| 25-29 | 25.52 | 47.15 | 28.78 | 21.85 | 19.10 | 17.86 | | | |
| 30-34 | 17.64 | 78.11 | 30.51 | 18.56 | 14.13 | 11.17 | 7.43 | .64 | |
| 35-39 | 10.46 | 45.81 | 30.12 | 17.45 | 10.67 | 7.42 | 3.87 | .36 | .22 |
| 40-44 | 5.82 | | 26.41 | 19.18 | 10.82 | 5.81 | 2.34 | .34 | .07 |
| 45-49 | 2.53 | | 17.32 | 17.31 | 10.20 | 5.07 | 1.64 | .20 | .01 |
| 50-54 | .89 | | (14.20) | 20.49 | 7.50 | 4.59 | 1.23 | .10 | .00 |
| 55 and over | .60 | | | 20.30 | 8.83 | 4.45 | 1.26 | .10 | .00 |

to 35-44 husbands. No similar resurgence of fertility appears according to age of wife. The 1948 data are not nearly as reliable because of the small population base, and the unusual fluctuation of the rates for under

[28] See footnote 11, 1940c and 1940d, above.

[29] See footnote 11, 1948a and 1948b, above.

20 wives cannot be wholly accepted. However, it may be noted from the 1948 results that as age of husband increases beyond 30, for wives aged 20-24, the fertility rate recovers; and a similar result appears in the next two age groups of wives having husbands with 10 or more years of seniority. In the first case (1940), older wives with husbands who are not too young show a more favorable birth performance with the older (but younger than they) husbands, in spite of the fact that the youngest husbands ordinarily show the highest rate; and in the second case (1948), a large seniority of the male tends also to raise the rate again. The data offer no explanation of the results, but we may suppose that the sexual component is a crucial factor. Youthful wives may be a revitalizing influence for older men, and youthful husbands (provided they are not too young) may have a similar effect for women over 35 years. However, a renewed desire for children by those in more mature years may also promote the same result.

In comparing the relative rates in 1940 the importance of youthfulness in a wife stands out clearly: husbands 55 years and over with wives under 25 years exceed the fertility rate for younger husbands who have wives 35-39 years; husbands 45-54 with wives under 25 have a slightly better performance than younger husbands over 25 years with wives aged 25 years or over; and, husbands 35-44 with wives under 20 easily surpass the performance of all younger husbands with wives aged 20-24. The 1948 results differ somewhat but similar comparisons can be found; for instance, husbands aged 35 to 44 have a higher performance rate with wives aged 20-24 than do younger husbands with wives aged 25-29 years.

The 1948 data show that, where both husband and wife were under 20, the fertility rate was lower than for the two neighboring groups, whereas according to the prevailing pattern the rate for this group should have been the highest. Allowance for illegitimacy would depress the rate for couples under 20 even further. In view of the statistical qualifications of the population sample for 1948 the result can only be regarded as suggestive, and cannot be taken as evidence of the biological hypothesis of adolescent sterility, recently reiterated at length by Ashley-Montagu.[30] If couples of this age group delayed their birth performance because of social reasons this too might effect the same result. It should

[30] Ashley-Montagu, M. F. Adolescent Sterility. *Springfield, Ill.*, C. Thomas, 1946.

also be noted that, even allowing for illegitimacy, the unilinear marital fertility rate for females in 1947 was highest for those aged 14-17 years.

## CONCLUSION

It should be remembered that these data do not show the occurrence of initial fertility, or the decline of biological fecundity with age. The height of the above rates in the earlier years is no doubt due to the occurrence of first births early in marriage, and the rapidity of decline in the rates is most likely due to the purposeful control of fertility, as families fill their social quota of births, and also as other conditions of life intervene.

Because of changes in the proportions of the population who are in the married state, changes in the pattern of fertility by age should be derived from calculations upon the married population, or the results may be quite misleading. A resurgence in fertility of married females at all age levels in this country since 1940, not to be wholly attributed to first births, may be due to favorable social and economic conditions of late. Whether this is actually a new trend marking the appearance of larger families, or merely a statistical illusion, remains to be seen. The change in the pattern of fertility, however, has not been as great as the customary indexes show.

The results of this study reveal that the rate of birth drops geometrically with increasing age of either party, and the joint ages of the pair interact to determine the actual performance. Fertility ceases in the woman around the age of 45, but the male procreative power wanes indefinitely onward. With advancing years the loss of fertility can be somewhat minimized, but not entirely counterbalanced, by the union with an increasingly younger spouse.

# THE POPULATION GROWTH OF FORMOSA*

BY SUNG-KEN QUO

## I. INTRODUCTION

IT IS well known to students of demography that the population growth on the island of Formosa is one of the most rapid in the various areas of the world. Its annual rate of natural increase in the year 1940 was 24.0 per 1,000 population. Some observers suggested that this island would become an example of the Malthusian theory, unless measures were taken to check the high birth rate, which was, for example, 43.7 live births per 1,000 population for the year 1940.

Most available data show, however, that a constant rate of increase in the population of a limited area never holds for any long period of time. On the contrary, the rate of population growth declines gradually with the flow of time, provided no cataclysmic alteration of the circumstances under which the population is growing has in the meantime intervened. The logistic curve discovered by Verhulst (1) and independently by Pearl and Reed (2) describes accurately the trend of the population growth in many confined areas. This curve enables not only the estimate of the populations of various countries in intercensal years, but also the prediction of future population growth (3).

According to the logistic law of growth, the population of any fixed area starts from a lower asymptote and grows slowly at first, but the rate constantly increases to a certain point where the rate of growth reaches a maximum. This point is the point of inflection of the population growth curve. After that point is passed, the rate of growth decreases progressively, till finally the curve stretches along nearly horizontally in close approach to the upper asymptote. The population cannot grow over the upper asymptote, unless some new factors or forces, which have never operated in the past, come into play to create a new cycle of population growth.

* From the Department of Biostatistics (Paper No. 267), School of Hygiene and Public Health, The Johns Hopkins University, Baltimore, Md.

It is, therefore, of great interest to investigate analytically to what extent Formosa follows the general law of human population growth and what are the prospects for the future population of this island.

## II. MATERIAL

The island of Formosa was ceded to Japan by China in 1895. After World War II the island was returned to China; all the Formosans

TABLE 1

*Formosan population, 1905-1942*

| YEAR | | POPULATION | YEAR | | POPULATION |
|---|---|---|---|---|---|
| Census | Intercensus | | Census | Intercensus | |
| 1905 | | 3 055 461 | 1925 | | 3 924 574 |
| | 1906 | 3 075 375 | | 1926 | 4 010 485 |
| | 1907 | 3 097 052 | | 1927 | 4 096 057 |
| | 1908 | 3 118 516 | | 1928 | 4 186 518 |
| | 1909 | 3 146 505 | | 1929 | 4 284 902 |
| 1910 | | 3 186 605 | 1930 | | 4 400 076 |
| | 1911 | 3 243 178 | | 1931 | 4 514 820 |
| | 1912 | 3 294 448 | | 1932 | 4 639 226 |
| | 1913 | 3 349 072 | | 1933 | 4 759 197 |
| | 1914 | 3 392 936 | | 1934 | 4 882 288 |
| 1915 | | 3 414 388 | 1935 | | 4 990 131 |
| | 1916 | 3 435 034 | | 1936 | 5 108 914 |
| | 1917 | 3 482 084 | | 1937 | 5 261 404 |
| | 1918 | 3 499 706 | | 1938 | 5 392 806 |
| | 1919 | 3 538 681 | | 1939 | 5 524 990 |
| 1920 | | 3 566 381 | 1940 | | 5 682 233 |
| | 1921 | 3 632 647 | | 1941 | 5 832 682 |
| | 1922 | 3 697 371 | | 1942 | 5 989 883 |
| | 1923 | 3 763 548 | | | |
| | 1924 | 3 827 112 | | | |

come of Chinese stock. During the Japanese regime, censuses of the population were taken at five-year intervals (4). The Formosan population enumerated in each of the eight censuses taken from the year 1905 to the year 1940 inclusive is given in Table 1. The population of each intercensal year, which is also included in the table, was calculated by adding to the population of the previous year the difference between the total births and the total deaths which occurred during that year.

As a matter of fact, emigration need not be taken into account in the case of the Formosans, who were confined to the island by the Japanese policy. They constituted from 95 to 98 per cent of the total population in Formosa during the Japanese regime. The Japanese and the tribal aborigines are excluded from the table.

The writer of this paper, who, as a native Formosan, knows well how strictly and accurately those censuses, as well as birth and death registrations, were taken during the Japanese regime, is fully convinced that the census data used in the present investigation are reliable. The eight census counts of population in Table 1 have been used for fitting the logistic curve.

### III. METHOD OF ANALYSIS

The equation of the logistic curve which can be fitted to the census data of the Formosan population is

$$y = d + \frac{k}{1 + e^{a+bx}} \tag{1}$$

where $y$ represents population, $x$ denotes time, and $d$, $k$, $a$ and $b$ are constants to be determined. Among these constants, $d$ is the lower asymptote which represents the total growth attained in the previous growth cycle, $d + k$ is the upper asymptote of the present cycle, $b$ the inherent rate constant, and $a$ the constant of integration dependent upon the position of the origin.

For the determination of these constants, we write equation (1) in the form

$$\log_e \frac{k - (y - d)}{y - d} = a + bx. \tag{2}$$

The common logarithm of $[k - (y - d)]/(y - d)$ is, therefore, a linear function of time, so that a straight line will be obtained when $[k - (y - d)]/(y - d)$ is plotted against $x$ on arithlog paper. From the curve of the population plotted against time on arithmetic paper, the upper and lower asymptotes were first estimated and $[k - (y - d)]/(y - d)$ was calculated for each census year by using the estimated values of $d$ and $k$. These values were plotted against $x$ on arithlog paper. If the values of the upper and lower asymptotes have been correctly estimated, this plot will be a straight line. If the values are too high or too low, the plot will curve up or downward, and we have to try new values for the upper or lower asymptote or both, till

$[k-(y-d)]/(y-d)$ will plot as a straight line on arithlog paper. Using the estimated values of $d$ and $k$, the values of $a$ and $b$ can be computed by the method of least squares.

The procedure stated above is but a first approximation of the four constants to the logistic curve. It is desirable, however, to obtain the best fitting curve by successive approximations which will correct the estimated values of all the constants and thus minimize $y$ deviations. Equation (1) is a function of $x$ which can be expanded by Taylor's theorem. Neglecting terms of higher order than the first, we have a linear expression for the correction terms which are to be determined. Now

$$f = d + \frac{k}{1 + e^{a+bx}},$$

and therefore

$$x_1 = \frac{\partial f}{\partial k_1} = \frac{Y_1 - d_1}{k_1},$$

$$x_2 = \frac{\partial f}{\partial a_1} = -\frac{(Y_1 - d_1)[k_1 - (Y_1 - d_1)]}{k_1},$$

$$x_3 = \frac{\partial f}{\partial b_1} = \frac{x(Y_1 - d_1)[k_1 - (Y_1 - d_1)]}{k_1},$$

$$x_4 = \frac{\partial f}{\partial d_1} = 1,$$

where $Y_1$, $d_1$, $k_1$, $a_1$ and $b_1$ are respectively the estimated values of $Y$, $d$, $k$, $a$ and $b$ from the first approximation.

Let $c_1 = k_2 - k_1$; $c_2 = a_2 - a_1$; $c_3 = b_2 - b_1$; $c_4 = d_2 - d_1$ in which $k_2$, $a_2$, $b_2$ and $d_2$ are the values of the constants that will be calculated from the second approximation. Then we fit the regression

$$Y = y' - Y_1 = c_1x_1 + c_2x_2 + c_3x_3 + c_4x_4$$

where $y'$ denotes the observed population. The values of the four correction terms $c_1$, $c_2$, $c_3$ and $c_4$ can be calculated by solving the following normal equations:

$$\begin{aligned}
c_1s_{11} + c_2s_{12} + c_3s_{13} + c_4s_{14} &= s_{1Y},\\
c_1s_{12} + c_2s_{22} + c_3s_{23} + c_4s_{24} &= s_{2Y},\\
c_1s_{13} + c_2s_{23} + c_3s_{33} + c_4s_{34} &= s_{3Y},\\
c_1s_{14} + c_2s_{24} + c_3s_{34} + c_4s_{44} &= s_{4Y}.
\end{aligned}$$

Here

$$s_{11} = \Sigma(x_1 - \bar{x}_1)^2;\ s_{12} = \Sigma(x_1 - \bar{x}_1)(x_2 - \bar{x}_2);$$
$$s_{1Y} = \Sigma(x_1 - \bar{x}_1)(Y - \bar{Y})$$

and so on.

If the second approximation to the best fitting curve thus obtained does not sufficiently minimize $y$ deviations, we may proceed to the third and the fourth approximations, and so forth, by repeating the same process, till the best fit to the curve is finally attained.

## IV. RESULTS

The final equation of the theoretical curve for the population growth of Formosa is

$$y = 2.833 + \frac{7.982}{1 + 49.0161e^{-0.4123x}} \qquad (3)$$

where $y$ denotes the population in millions, and $x$ time in base units of five years. In calculating the population at any particular year, the value to be substituted for $x$ is obtained from the formula $x = \frac{\text{Year} - 1900}{5}$, since the year 1905 was coded as 1 in the whole series of computation.

Table 2 presents the observed population at each census as compared with the population calculated from equation (3).

TABLE 2

*Formosan population as enumerated and as calculated from the logistic equation*

| CENSUS YEAR | POPULATION IN MILLIONS | | |
|---|---|---|---|
| | Observed | Calculated | Deviations (Obs.-Calc.) |
| 1905 | 3.055 | 3.0716 | — 0.0166 |
| 1910 | 3.187 | 3.1880 | — 0.0010 |
| 1915 | 3.414 | 3.3573 | + 0.0567 |
| 1920 | 3.566 | 3.5992 | — 0.0332 |
| 1925 | 3.925 | 3.9362 | — 0.0112 |
| 1930 | 4.400 | 4.3896 | + 0.0104 |
| 1935 | 4.990 | 4.9712 | + 0.0188 |
| 1940 | 5.682 | 5.6741 | + 0.0079 |

The standard deviation of the observed values about the curve is given by:

$$\sigma_{y.x} = \sqrt{\frac{\Sigma(\text{dev.})^2}{\text{Degrees of freedom}}} = \sqrt{\frac{0.005243}{8 - 4}} = \pm 0.036.$$

It is evident that equation (3) furnishes an excellent approximation to the course of the known changes in the population of Formosa.

Fig. 1 represents the theoretical curve of the population growth in Formosa; the small circles indicate observed population. It is clear that there is extremely close agreement between observation and theory. A better fit could hardly be expected. The curve is a typical logistic curve, symmetrical at its point of inflection, which is located exactly half-way between the two asymptotes. At this point there is a maximum

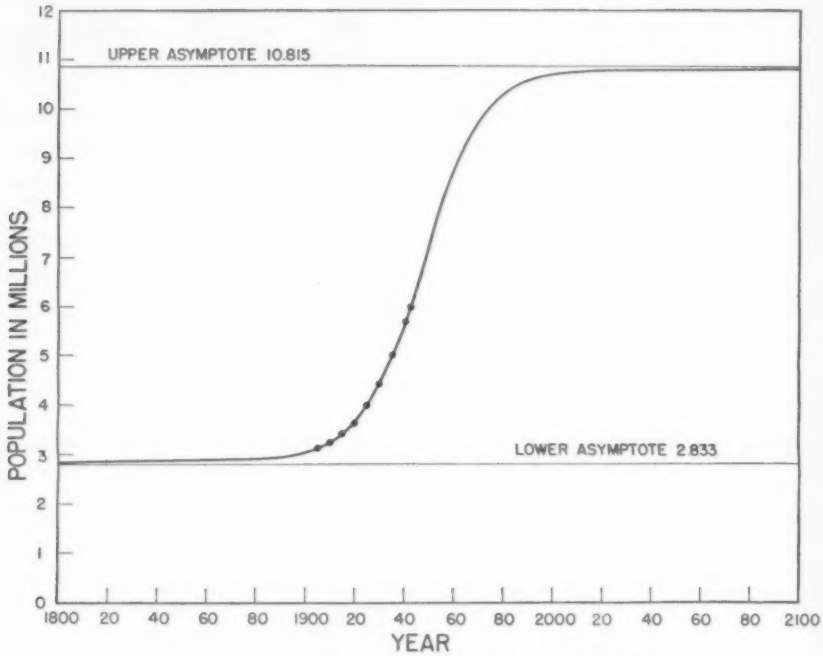


FIG. 1. THE POPULATION GROWTH OF FORMOSA

rate of population growth. The date at which the maximum growth rate occurs can be easily determined either from the curve or from equation (3).

The growth curve of the Formosan population starts from a lower asymptote of roughly 2.833 millions which was fairly closely approached in the year 1800. The population continues to grow at an increasing rate up to the point of inflection at which the maximum rate of growth is attained and from which the growth rate begins to decrease steadily, till the population becomes stationary from the year 2000. The calcu-

lated point of inflection in the present growth cycle occurred about August 1946 when $y = k/2$. The upper limit, which is 10.815 millions, will be reached at about the year 2100. This will give the island (area, 13,890 square miles) a maximum density of population of 778.6 persons per square mile or 300.7 persons per square kilometer.

## V. DISCUSSION

It is obvious that the logistic curve gives an excellent fit to the population growth of Formosa. Therefore the probable population at any particular year may be calculated from the equation of the fitted curve obtained. The lower asymptote of 2.833 millions indicates a prior growth cycle which brought the population up to about that figure. This suggests that the present cycle of population growth started probably sometime about the year 1800 from the top or near the top of the prior growth cycle. The point of inflection for the present growth cycle is already passed, in view of the fact that the annual rate of increase in the Formosan population tends to decline since a few years ago. In other words, Formosa has already passed its period of most rapid population growth in the present cycle.

The basic postulate of the logistic theory of population growth is that any particular population can be expected to continue to follow in its later growth the same logistic curve which it has followed in its earlier growth, so long as there is no serious or cataclysmic alteration of the conditions, climatic, geographical, biological, economic or social, under which the population has been growing. Therefore, predictions based on the extrapolated portion of the growth curve should be made on this assumption.

The predicted upper limit, i. e. the point of saturation, of the Formosan population is 10.815 millions, which will be reached at about the year 2100. Obviously, this prediction is based upon the assumption that no other factors will operate to alter the course of the present curve; moreover it should be noted that the absolute size of the population at the asymptote is subject to a large probable error, so that the above figure is at best only a rough approximation.

Formosa is at the present time the only area of China where statistical data on population are available. We have reason to believe that these data are sufficiently accurate to be used for the present investigation. The tribal aborigines are living scattered in the mountains and have no economic relationship with the Formosans coming of Chinese stock.

The Japanese who have been repatriated from the island are replaced by the Chinese coming from China mainland. It seems probable, therefore, that there will be no violent change of conditions which may affect the normal growth of the Formosan poplation. It is possible that the population, which has just passed its point of maximum growth, will hereafter follow the later portion of its logistic growth at a gradually declining rate.

However, if some new factor or force intervenes in the future to alter seriously the conditions under which the population is now growing, the results would be merely to throw the growth trend of the population to a new curve of the same type as the present one, but having a different rate.

## VI. SUMMARY

In order to study the growth trend of the population on the island of Formosa, eight successive census counts were fitted with the logistic curve, which was found to give an excellent fit to the observed population. The values for the upper and lower asymptotes and the point of inflection, as well as the years in which these values were or will be attained, have been calculated from the logistic equation obtained, and the growth trend of the Formosan population has been discussed.

As a matter of course, the predictions have been made under the assumption that no new factors or forces will come into play in the future to alter seriously the conditions under which this population has been growing.

Acknowledgments are gratefully accorded by the author to Professor W. G. Cochran and Drs. C. P. Winsor and R. V. Rider for their helpful suggestions in the preparation of this paper.

## BIBLIOGRAPHY

I. (a) VERHULST, P. F., Notice sur la loi que la population suit dans son accroissement. *Correspondance mathématique et physique publiée par A. Quetelet*, X, pp. 113-121, 1838.

(b) VERHULST, P. F., Recherches mathématiques sur la loi d'accroissement de la population. *Nouveaux mémoires de l'Académie Royale des Sciences et Belles-Lettres de Bruxelles*, XVIII, pp. 1-38, 1845.

(c) VERHULST, P. F., Deuxième mémoire sur la loi d'accroissement de la population. *Ibid.*, XX, pp. 1-32, 1847.

2. (a) PEARL, R. and REED, L. J., On the rate of growth of the population of the United States since 1790 and its mathematical representation. *Proc. Nat. Acad. Sci.*, Vol. 6, pp. 275-288, 1920.

(b) PEARL, R. and REED, L. J., A further note on the mathematical theory of population growth. *Proc. Nat. Acad. Sci.*, Vol. 8, pp. 365-368, 1922.

(c) PEARL, R. and REED, L. J., On the mathematical theory of population growth. *Metron*, Vol. 3, No. 1, pp. 6-19, 1923.

3. (a) PEARL, R. and REED, L. J., Predicted growth of population of New York and its environs. Plan of New York and Its Environs, New York, No. 4, pp. 1-42, 1923.

(b) PEARL, R., Studies in Human Biology, Chapters XXIV and XXV. Williams and Wilkins Company (*Baltimore*), 1924.

(c) PEARL, R., The curve of population growth. *Proc. Amer. Phil. Soc.*, Vol. 63, pp. 10-17, 1924.

(e) PEARL, R. and REED, L. J., The logistic curve and the census count of 1930. *Science*, Vol. 72, pp. 399-401, 1930.

(e) PEARL, R., REED, L. J. and KISH, J. F., The logistic curve and the census count of 1940. *Science*, Vol. 92, pp. 486-488, 1940.

4. (a) GOVERNMENT-GENERAL OF FORMOSA, Statistical Annual of Formosa. *Taihoku, Formosa*, 1905-1942.

(b) GOVERNMENT-GENERAL OF FORMOSA, Vital Statistics Yearbook of Formosa. *Taihoku, Formosa*, 1905-1942.

(c) OFFICE OF THE CHIEF ADMINISTRATOR, PROVINCE OF TAIWAN, BUREAU OF STATISTICS, Statistical Abstract of the Province of Taiwan for the Past Fifty-One Years. *Taipeh, Formosa*, 1946.

# INDEX TO VOLUME XXII

(Asterisked entries are titles.)


Age(s), of couples, and marriage fertility, 281
visual acuity and color vision, 1
*Analysis of heart measurements of growing boys, an, 271
*Application of gene frequency analysis to the interpretation of data from twins, an, 136

Behrens-Fisher problem, 125
Behrens (Reed-Muench) method for LD 50, sampling error of, 151
Bioassay, sampling variation in, 151
Births, frequency by day of week, 218
BLISS, C. I., and YOUNG, M. S., An Analysis of Heart Measurements of Growing Boys, 271-280
Blood group estimation, statistical analysis of, 191

CHAPANIS, ALPHONSE, Relationships between Age, Visual Acuity and Color Vision, 1-33
Color vision, visual acuity and age, 1
Control group, graphic method of comparing with experimental group, 65
COOK, ELLSWORTH B., and WHERRY, ROBERT J., The Urinary 17-Ketosteroid Output of Naval Submarine Enlisted Candidates during Two Stressful Situations, 104-124
Correlation models, and semantics, 75

Day of week and frequency of births, 218
Definition of group to be followed, in follow-up studies, 233
DENSEN, PAUL M., Long-time Follow-up in Morbidity Studies: The Definition of the Group to be Followed, 233-237
Dermatoglyphics, and race, 34
Digital patterns, and race, 34
DORN, HAROLD F., Methods of Analysis for Follow-up Studies, 238-248

Factor analysis of ketosteroid output, 104
Fertility by ages of couples, 281
Fifty per cent end-point, sampling variation of, 151
Follow-up studies, definition of group, 233
methods of analysis, 238
theory, 249
Formosa, population growth, 293

Gene frequency analysis applied to twin data, 136
*Generalization of the Behrens-Fisher problem, on a, 125
*Graphic method for comparing percentages and means of control and experimental groups, a, 65

Growth, and temperature, 71
of boys, and heart measurements, 271
sexual, of Negro and white boys, 146
GRUBER, JACOB W., The Science of St. George Mivart, 222-232

Handedness among twins and immediate relatives, 136
HARRIS, DALE B., A Note on the Relationship of Births to Day of Week, 218-221
HARRIS, T. E., MEIER, PAUL, and TUKEY, JOHN W., Timing of the Distribution of Events between Observations: A Contribution to the Theory of Follow-up Studies, 249-270
Heart measurements of growing boys, 271
Height-weight trends, and temperature, 71
HOGBEN, LANCELOT, and KEMP, KENNETH W., Statistical Models Bearing on the Semantics of Correlation. III. Expectation-Accomplishment Models, 75-103.

Kemp, Kenneth W., *see* HOGBEN, LANCELOT
Ketosteroid output, and stress, 104

LD 50, sampling variation of, 151
Logistic curve, describing population growth of Formosa, 293
*Long-time follow-up in morbidity studies: the definition of the group to be followed, 233
Losses, in follow-up studies, 238, 249

*Marriage fertility by ages of couples, 281
Maximum likelihood estimates in follow-up studies, 249
of blood group frequencies, 191
Means, graphic method for comparing, 65
Measurements, heart, of growing boys, 271
Meier, Paul, *see* HARRIS, T. E.
Menarcheal age, trends, and temperature, 71
*Methods of analysis for follow-up studies, 238
MILLS, CLARENCE A., Temperature Influence over Human Growth and Development, 71-74
Mivart, St. George, 222
MONAHAN, THOMAS P., Marriage Fertility by Ages of Couples, 281-292
Morbidity studies, follow-up, 233, 238, 249

Percentages, graphic method for comparing, 65
*Phylogeny of digital patterns in man and its bearing on racial affinities: a study in human ecology, 34
PIZZI, MARIO, Sampling Variation of the Fifty Per Cent End-point, Determined by the Reed-Muench (Behrens) Method, 151-190.
*Population growth of Formosa, the, 293
Probability ellipses, approximate construction for, 191
Probit method of determining sampling variation of LD 50, 151
Psychological stress, and ketosteroid output, 104

QUO, SUNG-KEN, The Population Growth of Formosa, 293-301

Race, and digital patterns, 34
RAMSEY, GLENN V., Sexual Growth of Negro and White Boys, 146-149.
Reed-Muench method of determining sampling variation of LD 50, 151
*Relationship(s) between age, visual acuity and color vision, 1
* of births to day of week, a note on the, 218
RIFE, DAVID C., An Application of Gene Frequency Analysis to the Interpretation of Data from Twins, 136-145

Sampling variation in bioassay, 151
* of the fifty per cent end-point, determined by the Reed-Muench (Behrens) method, 151
SCHREK, ROBERT, A Graphic Method for Comparing Percentages and Means of Control and Experimental Groups, 65-70
*Science of St. George Mivart, the, 222
Sexual growth of boys, Negro and white, 146
*Statistical analysis of the A-B-O blood groups, 191
models bearing on the semantics of correlation. III. Expectation-accomplishment models, 75
STEVENS, W. L., Statistical Analysis of the A-B-O Blood Groups, 191-217
Stress and ketosteroid output, 104

Taylor, Griffith, Zones and Strata Theory of, 34
*Temperature influence over human growth and development, 71
*Timing of the distribution of events between observations: a contribution to the theory of follow-up studies, 249
Tukey, John W., *see* HARRIS, T. E.
Twin data, gene frequency analysis of, 136

*Urinary 17-ketosteroid output of naval submarine enlisted candidates during two stressful situations, the, 104

Variance, analysis of, 75
partition of, 75
Visual acuity, age, and color vision, 1

WALSH, JOHN E., On a Generalization of the Behrens-Fisher Problem, 125-135
Weight of boys, and heart size, 271
-height trends and temperature, 71
Wherry, Robert J., *see* COOK, ELLSWORTH B.
Withdrawal rates, in follow-up studies, 238
WOLFE, ROY I., The Phylogeny of Digital Patterns in Man and Its Bearing on Racial Affinities: A Study in Human Ecology, 34-64

Young, M. S., *see* BLISS, C. I.

Zones and Strata Theory of Griffith Taylor, 34

MEDICAL LIBRARY

# HUMAN BIOLOGY

## a record of research

DECEMBER, 1950

UNIVERSITY
OF MICHIGAN
FEB 26 1951
MEDICAL
LIBRARY

Vol. 22 No. 4

PUBLISHED QUARTERLY BY
THE JOHNS HOPKINS PRESS, BALTIMORE, MARYLAND, U.S.A.

# HUMAN BIOLOGY

## a record of research

FOUNDED BY RAYMOND PEARL

*Edited by*

CHARLES P. WINSOR

*in association with*

A. W. FREEMAN PHILIP E. SARTWELL ADOLF MEYER

CURT P. RICHTER ADOLPH H. SCHULTZ

*Managing Editor*

FRANCES D. NICOLLS

HUMAN BIOLOGY is published quarterly by The Johns Hopkins Press, Baltimore, Maryland, U. S. A. The advance subscription price is $5.00 a year (foreign postage 35 cents; Canadian 13 cents). Single current issues are $2.00 each; back numbers, $3.00 each; back volumes, $6.00; plus postage. Subscriptions for less than a year are charged at the single issue price.

Subscribers should notify the publishers of change in address at least two weeks before publication of the issue with which change is to take effect. No claim for non-receipt of an issue will be entertained unless made immediately upon receipt of the next succeeding issue.

Manuscripts and communications regarding editorial matters should be addressed to Editor, Human Biology, 615 N. Wolfe St., Baltimore—5, Maryland, U. S. A.

THE JOHNS HOPKINS PRESS, *Publishers*

BALTIMORE—18, MARYLAND, U. S. A.

Entered as Second Class matter, February 18, 1929, at the Post Office at Baltimore, Md., under the Act of March 3, 1879.


Copyright 1950 by The Johns Hopkins Press